Semi-supervised Clustering of Visual Signatures of Artworks

A human-in-the-loop approach to tracing visual pattern propagation in art history using deep computer vision methods.

Master Thesis
Submitted on 1st July, 2022
Student: Ludovica Schaerf
Supervisor: Frédéric Kaplan
Advisor: Paul Guhennec
External expert: Bernard Aikema
Lab: Digital Humanities Lab (DHlab)
École Polytechnique Fédérale de Lausanne
Lausanne, EPFL, 2022

“Il qual disegno non può avere buon’origine,
se non s’ha dato continuamente opera a ritrarre cose naturali,
e studiato pitture d’eccellenti maestri,
e di statue antiche di rilievo,
come s’è tante volte detto”
(Vasari, [1550] 1876II, Vite de’ più eccellenti pittori, scultori e architettori, Chapter 15)

“which artwork cannot have a respectable origin,
if one has not industriously depicted the natural things,
and studied the paintings of the excellent masters,
and the illustrious statues of antiquity,
as we have repeated already”
(My own translation of: Vasari, [1550] 1876II, Vite de’ più eccellenti pittori, scultori e architettori, Chapter 15)

Acknowledgements

When I was 12, my middle school teacher, chatting with my mum, told her she had understood a great quality of mine: to be able to channel resources (in Italian: 'convogliare risorse'). Of course, at the time that meant that I managed to involve my whole family team in doing my homework: my grandma for Literature, my dad for Maths, my grandpa for Physics and my mum for Arts, with some flexibility.
While I have evolved to more subtle approaches to channelling resources, I would like to dedicate this thesis to all the people that make me and my work what it is.
In this, I would like to thank all my supervisors and assistants. Paul, for his pain au chocolate, the illuminating discussions and encouragement to dear. Prof. Kaplan, for all the opportunities, the directions, the scientific impetus. Dr. Di Lenardo, for her explanations on the art historical subject matter of this thesis, and for the very practical help in drafting the Introduction.
I would like to thank my parents, who were even up late the day before the deadline offering to help in any way they could, for always being supportive and participative. My brother for constantly bringing up new sides of me and allowing me to do the same with him. My grandparents, who are my source of inspiration and aspiration. Davide, my number one supervisor, my daily exchange of thoughts, of emotions, of experiences.
I would like to thank my art history team, my 'sister since middle school' Eleonora and her friend Ludovico for the coffee talks on patterns in art history, and Ravi for helping me take the first steps in the annotation platform and being my mate in the lab.
I cannot forget to thank the labs in the Digital Humanities section, from the Musicology to the History lab, with whom I started to feel a sense of belonging to a community of researchers, and not an unproblematic one, rather the best kind, one of those that are in the course of finding their balance. In particular, I think of Gabriele, Yannis, Yuanhui, and Haaeun.
Finally, I extend a huge thank you to all the invaluable people in my life, from Rome to Amsterdam and to Lausanne.

Lausanne, 1 July 2022 L.S.

Abstract

The project Replica, about six years ago, paved the way to computational studies of visual patterns in art history. Simultaneously, it created the possibility for art historians to trace the propagation of patterns throughout the history of art. During the project, an image retrieval network was set up to discover artistic patterns given an input image. Despite successfully serving monographic needs and targeted search attempts, the network does not propose spontaneous discoveries. In this thesis, we eliminate the middle man of the input image, creating clusters of artworks sharing a common pattern propagation. The clusters are integrated further with a 2D coordinate-based visualisation, which provides an organic view of the evolution of the patterns in art history.
In this effort, we demonstrate the effectiveness of fine-tuning deep learning models on a set of visual connections using a compound Hinge loss and ResNeXt architecture. Moreover, we show that clustering the trained visual signatures with OPTICS yields remarkable precision. We emphasise the importance of the semi-supervised learning of the clusters, proving the qualitative and quantitative improvement over generic clustering methods. Furthermore, we close the loop of the semi-supervised clustering through the annotation of the new findings in the clusters proposed, and retraining thereof. In total, we add over 700 new images to the set of slightly over 1800 existing visual connections. We find, in addition, examples of cross-domain, architectural, design and sketch based patterns, which were previously outside the scope of the known visual connections.

Keywords: visual patterns, semi-supervised, clustering, triplet learning, digital art history

Code:
https://github.com/ludovicaschaerf/Cini_TDA - Main
https://github.com/ludovicaschaerf/pattern_clusters
https://github.com/ludovicaschaerf/pixplot_pattern_propagation

Website: https://patternclusters-api-heroku.herokuapp.com/

Résumé

Le projet Replica, il y a environ six ans, a ouvert la voie aux études informatiques des motifs visuels dans l'histoire de l'art, ouvrant par la même occasion aux historiens de l'art la possibilité de retracer à grande échelle la propagation des motifs dans l'histoire de l'art. Pour ce faire, un réseau neuronal de recherche d'images a été mis en place afin de découvrir des motifs artistiques à partir d'une image d'entrée. Bien qu'il ait répondu avec succès aux besoins monographiques et aux tentatives de recherche ciblées (c'est-à-dire au cas par cas), le réseau ne propose pas de découvertes spontanées. Cette thèse répond à ce besoin et propose l'élimination de l'intermédiaire de l'image d'entrée, en créant des clusters d'œuvres d'art partageant une propagation de motif commune. Les clusters sont ensuite intégrés à une visualisation 2D, qui suggère une vue organique de l'évolution des motifs dans l'histoire de l'art.

Dans cet effort, nous démontrons l'efficacité du *fine-tuning* des modèles d'apprentissage profond sur un ensemble de connexions visuelles en utilisant une perte de charnière composée et une architecture ResNeXt. De plus, nous montrons que le regroupement des signatures visuelles entraînées avec OPTICS donne une précision remarquable. Nous soulignons l'importance de l'apprentissage semi-supervisé des clusters, en prouvant l'amélioration qualitative et quantitative par rapport aux méthodes de clustering génériques.

De plus, nous bouclons la boucle du clustering semi-supervisé par l'annotation de nouvelles découvertes dans les clusters proposés, et leur ré-entraînement. Au total, plus de 700 nouvelles images ont été ajoutées à l'ensemble d'un peu plus de 1800 connexions visuelles déjà existantes. Nous trouvons, en outre, des exemples de modèles inter-domaines, architecturaux, de conception et de croquis, qui étaient auparavant hors du champ des connexions visuelles connues.

Mots-clés: motifs visuels, semi-supervisé, clustering, apprentissage par triplet, histoire de l'art numérique

Contents

List of Figures

List of Tables

1 Introduction

1.1 Preamble

In the Introduction to this thesis, as well as in other parts of the work, we introduce some concepts regarding the art historical discussion on patterns. Given the lack of an art historical training, we wish to present these concepts as mere thoughts which, we believe, may be of help in framing the general discussion. We deem these considerations by no means exhaustive for an art historical discussion.

1.2 The transmission of visual patterns as a distinctive aspect of artistic practice

The history of art is not only about invention, but also about a continuous improving and reworking of existing scenes (Gombrich, 1965; Vasari, [1550] 1876II). According to some thinkers, art is a continuous progress towards the most natural representation of reality (Gombrich, 1961). In this quest for naturalism, the artist is forced to begin from schemata inherited by others to represent the world, as it cannot be grasped through the human senses. In Vasari, [1550] 1876II, for an artwork to be of great value, the artist must have depicted the natural things, studied the paintings of the excellent masters and the ancient statues with meticulous repetition.

In fact, for Vasari, the stages of confrontation with the artistic tradition of the models of the past pass through the *traslatio*, i.e. the simple transfer of a visual model aimed at reproducing with the main purpose of making a source available to be copied again; the *imitatio*, characterised by a direct but not passive quotation enriched, thus, with some variation; and the *aemulatio*, i.e. a surpassing of the original to create a new work that mixes several important quotations and demonstrates the author's artistic knowledge and ability to reinvent the models with a new composition (Muller, 1982; Warners, 1956). A particular role is played by *inventio*, as a peculiar characteristic of certain artistic individualities capable of going beyond established tradition and models and introducing new elements in this continued circulation of visual motifs

(Jost, 1966; Loh, 2007). Already the treatise writers of the modern era were clear about this distinction between copied or replicated elements and the original, understood as archetypo, a kind of Ur-type that represented the idea, i.e. the fruit of an original creation characterised by its originality, without, apparently, being indebted to copying anyone else (Loh, 2007; Spear, 2002).

To envisage art in the Vasarian optic helps, among others, to see the value of repetition, which orients the learning and the progress. In fact, artistic historiography has well highlighted how the repetition, the copy of details or of whole compositions, the propagation of visual motifs, constitutes one of the distinctive aspects of artistic practice, in particular for European painting of the modern age (Gombrich, 1961; Wittkower, 1965).

This imitation of visual models of the past is a fundamental concept of artistic pedagogy, well described by Renaissance critics and structured in the late modern era through the establishment of the Academies (Muller, 1982; Vander Auwera, 2007). In the rhetoric of artistic creation, be it literary or pictorial, the replication of visual motifs within the same artistic atelier or as a direct quotation between artists, even with a significant temporal and chronological distance, is considered a valuable element, demonstrating the awareness of the artist who produces it and whose path was inspired by the great artists of the past and in continuous comparison with them (Dempsey, 1980).

To the local propagation inside a workshop, we add the idealism of the antiques, which we believe is manifested in the early XVI century fascination for Roman and Greek antiques, the repeated copy making of the Lacoonte[I] and elements of the Domus Aurea[II] (Vasari, [1550] 1876II). This idealism concurs to create a second playground for artist, that eventually acquire the pictorial repertoire of the antiques. This idealism is recovered in Warburg's *Nachleben der Antiken*. In his work, Warburg emphasises how such ancient patterns reappear in the form of intrusions, of motifs propagated from the past more or less consciously (Didi-Huberman, 1996; Diers et al., 1995.

The propagation of motifs in Warburg, however, is a rather organic, all-encompassing, phenomenon, bridging the fields of psychology, sociology, cultural studies (Didi-Huberman, 1996). Our consideration of patterns is more delimited than Warburg's, as we do not draw continuities between mere cultural affinities unless they are expressed visually in the artwork. In this optic, we find our definition of patterns to be more akin to Panofsky, 2018's pre-iconographic level, of recognition of the natural subject matter. In fact, a pattern, although often times tightly linked to its iconological meaning, must be visible at the bare visual level of shapes and forms.

Beyond the pedagogical and idealists forces of patterns that we mentioned thus far, the reasons that surround the socio-economic sphere should not be overlooked. A field where the analysis

[I] The discovery happened in Rome in 1506 and the marble group immediately received large attention Martin, 1968

[II] The date of the discovery of Nero's Domus Aurea is uncertain, it can be collocated towards the end of the XV century La Malfa, 2000

Figure 1.1: Identification of pattern transmission in Titian's works and workshop Image compiled on the basis of the study Le Botteghe Di Tiziano, Florence, 2009.

of the propagation of visual patterns proves to be significant is that of the history of collecting and the geography of art because some of these motifs possess a considerable artistic fortune, that implies a considerable geographical propagation (Kaufmann, 2004). In fact, we can see patterns as reflexive of the cultural sphere of the taste of the centre, dictating the dominant style and compositions and their spread (Kaufmann et al., 2015).

The socio-economic and cultural sphere is guided by interest, that, in turn, is stimulated and stimulates the production of patterns. In this setting, the role played by patrons was frequently at the heart of pattern propagation. We enumerate among the examples, cases where, for instance, after seeing an impressive artwork, patrons dispense commissions, to the author of the original work or to copyists, to replicate the piece. A case such as Ghirlandaio's Coronation of the Virgin near Narni that was copied under patrons' orders by La Spagna and, successively, Siculo (Castelnuovo et al., 2009). Oppositely, it may arise from a situation of rejection by that patron, who commissioned a work which ultimately did not meet their favour. That piece would end up being sold elsewhere and a similar but revised work being produced for the patron, as in the case of Caravaggio's conversion of Saint Paul (Racco, 2016). Cases such as Neri di Bicci's prolific serial production for the art market, or that of the Flemish artists, whose workshops realise a mass series of works to be sold in the primary market, are important instances of the economic sphere of patterns (Etro and Pagani, 2013). In the production of the workshop, of copyists, and numerous others, we observe the pedagogical, idealist and socio-economic spheres coinciding and coexisting.

From the point of view of the methodology of art-historical research and the importance of studies on pattern, the analysis of the transmission of motifs is a fundamental aspect to be

Figure 1.2: Incoronazione della Vergine. Retrieved from here. On the left, Ghirlandaio's original Incoronazione della Vergine near Narni. In the centre, Giovanni di Petro, called Le Spagna's version, 1511, Trevi. On the right, Giacomo Santori di Jacopo Siculo's version, Norcia (Castelnuovo et al., 2009).

considered in forming an artist's catalogue, discarding or including works that possess certain clues of authorship (Patrick, 2014). The analysis of visual specificity in motifs, is also used to identify originals and fakes (Guichard, 2010).

Historiography has also highlighted how the analysis of visual motives makes it possible to clarify the hierarchical relationships between the painters involved in the same atelier, between the master and his pupils for example, to the point of establishing which work can be included or excluded from the artistic ecosystem of an author (Dal Pozzolo, 2006; Tagliaferro et al., 2009).

It follows that in order to study the reasons for the collector's fortune of a certain work, one must necessarily analyse from a stylistic and compositional point of view the fortune of a certain visual motif, be it a simple detail or an entire composition. Now, a systematic reconnaissance of historiography essentially reveals two critical aspects, which are in fact the reasons for this thesis. The first aspect concerns the shift, from a qualitative and specific plane, the visual motif, to a more generic, schematic and computational one, the visual pattern, and the textual or visual definition of this concept. The second aspect, partly a consequence of the first, thematises the methodological difficulty ascertained so far in confronting a large-scale comparison of the propagation of such patterns.

With regard to the first issue, we have decided not to pre-empt the conceptualisation of what a visual pattern might be as it is discussed by historiography, but consider some specific studies on visual transmission among authors of the modern era following precisely that mechanism of translation, imitation translation, and direct quotation with copies (di Lenardo et al., 2016; B. Seguin et al., 2016; B. L. A. Seguin, 2018).

Figure 1.3: Caravaggio. Conversione di San Paolo. Retrieved from here. On the left, the first version of the Conversion of Paul. The light of God has blinded Paul who fell from the horse in a chaotic and distressing scene. On the right, the second and more traditional version of the scene, with a calm open-armed Paul (Racco, 2016)

We propose thus to consider patterns as *purely visual repetitions that range from exact copies to clear visual inspirations (of a part or the whole artwork) caused by concurring factors of pedagogy, idealism and socio-economic and cultural reasons and manifesting a clear chain of observation.* By clear chain of observation we mean the presence of self-manifest evidence that for the artist to have replicated a pattern, they must have seen (directly or through other patterns) the original work. This involves that the second artwork cannot possibly be so similar to the first if it were not for direct or mediated observation of the first, or an intermediary artwork featuring the same pattern.

Patterns can be therefore summarised as *visual clues of transmission* as we assume that the same visual pattern cannot be invented independently by different authors. This, in return, constraints to consider only patterns sufficiently specific to be confident a multiple origin is impossible.

The second aspect from which this thesis stems is the methodological constraint of the scalability of research. So far, when historiography deals with patterns, it takes into consideration an author, his workshop, his circle, more rarely the greatest masters in comparison with each other (Titian, Rubens i.e.) but hardly measures itself with a catalogue extended in chronology and granularity that also includes minor artists, for example, undoubtedly fundamental links in the pattern transmission chain. Visual research of these patterns with traditional methods is almost impossible to manipulate due to the quantity of visual sources of comparison. In fact, it theoretically requires to be able to compare hundreds of thousands of artworks concurrently. Because this involves scanning billions of pairs of artworks, it is impossible to fully map the patterns in art history manually. Additionally, already existing textual searches in artistic datasets cannot detect the full range of patterns and are inherently limited to a single

figurative case.

With computational methods, the possibility of detecting large scale pattern propagation was opened up, creating the opportunity to endeavour in such comparison. Given the immense potential of this approach, this study aims to structure a computational approach to identify, extract and classify these similarities in a quantitatively consequential corpus, thus enabling the articulation of a cartography of propagation that is undoubtedly broader and more scalable than that manageable with traditional methodologies. Recent advances in digital humanities for art history lay the ground for developing the task.

1.3 The efforts in digital art history

In the last decade, the field of Computer Vision has witnessed significant technical development (Goodfellow et al., 2016). This has prompted researchers to investigate possible applications of the new techniques to visual arts. A special interest has been devoted to the classification of paintings based on learned (semantic) features. The first wave of this research was mainly circumscribed to the classification of artist, style, genre and year of paintings datasets. This does not come as a surprisingly, as these metadata fields are available in almost all published artwork datasets (such as the Rijksmuseum, Wikiart, Tate, MoMA datasets).

Already in the first decade of the century, the earliest attempts were made at classifying paintings. These first investigations, however, only dealt with features reflecting basic visual properties of a computer image, including colour, gradient orientation, pixel intensity, edges. Around the year of 2014, with the wave of success of deep computer vision techniques, researchers suddenly drifted away from feature engineering and moved towards feature learning (Goodfellow et al., 2016).

One of these first attempts was the artist attribution experiment using the PigeoNet architecture, a neural network comprising five CNN layers and three FNNs (Feed-forward Neural Networks, Van Noord et al., 2015). The network was trained on the Rijksmuseum public dataset and obtained a mean class accuracy of roughly 0.76 for the 100 artists (Van Noord et al., 2015). Although the results were comparable to those achieved by low-level features, these semantic level features showed more promise as their performance was proved to increase faster with larger datasets.

With the success of ResNet and other object recognition networks, researchers began to include pre-trained weights at the top of their networks, in order to transfer the knowledge learned by these weights to their task (K. He et al., 2016). One of the first instances of this trend is an investigation by Lecoutre et al., 2017, who used AlexNet and ResNet trained on ImageNet with 20 layers of fine-tuning to classify paintings into their style. The use of transfer learning and deep fine-tuning allowed, respectively, to overcome the issue of the sparsity of the data and of heterogeneity of tasks, obtaining an overall best accuracy of 0.62 with ResNet, over 25 classes. Along the same line, Sabatelli et al., 2019 conducted a more comprehensive evaluation

of different pre-trained models for painting classification. In fact, they used four pre-trained architectures (VGG19, Inception-V3, Xception and ResNet50) and assessed their performance on three different tasks (attribution of author, material and artistic category), both fine-tuning and not. ResNet50 outperformed all other architectures after fine-tuning, while simpler networks, like VGG19, had the highest off-the-shelf accuracy. Cetinic et al., 2018 conducted a comprehensive fine-tuning experiment on CNNs on five classification tasks (author, genre, style, time-frame, and nationality recognition), three dataset (Web Gallery of Art, WikiArt and TICC) and three weight initialisations (scene recognition, sentiment prediction and object recognition). They obtained the best performance with scene recognition, sentiment prediction initialisations and showed the importance of fine-tuning in tasks with numerous and imbalanced classes.

Departing from the well-investigated tasks, Deng et al., 2021 proposed a method to quantitatively assess the representativity of an artwork in the body of work of each artist. They use a style-enhanced deep model trained on artists recognition, selecting their most representative paintings. The authors obtain an accuracy above 0.7 with the ResNet architecture. In 2018, Elgammal et al., 2018 analyse the features extracted from style recognition in correlation to Wölfflin's concepts in art history[III]. To be able to interpret the results, the model results was reduced to a low number of dimensions, using principal component analysis (PCA). The paper showed that 10 modes contain 95% of the variation, and, retaining the 3 modes of highest variation, the correlation (Pearson correlation coefficient) with time and with a quantification of Wölfflin artistic concepts is high (Elgammal et al., 2018).

Paper	Year	Task	Contribution
Van Noord et al., 2015	2015	Artist classification	Learned descriptors
Lecoutre et al., 2017	2017	Style classification	Fine-tuning
Sabatelli et al., 2019	2019	Multiple classification	Architecture comparison
Cetinic et al., 2018	2018	Multiple classification	Task pre-training
Deng et al., 2021	2021	Representativity	Original task
Elgammal et al., 2018	2018	Feature-concept correlation	Interpretation of results
Saleh et al., 2016b	2016	Artist influences	Time awake influence detection
Castellano et al., 2021	2021	Artist influences	Simple pipeline
B. Seguin et al., 2016	2016	Pattern detection	Triplet learning
Shen et al., 2019	2019	Pattern detection	Self-supervision

Table 1.1: Overview table of the Digital Visual Art efforts.

Most akin methodologically to our investigations, there have been many attempts in the recent years to trace patterns or, more broadly, influences in art. Some efforts worth mentioning include Saleh et al., 2016a, who addressed the problem of automatically finding influences

[III] The concepts have been introduced by Wölfflin in 1915 in his famous book 'Fundamental Concepts of Art History', where he delineates 5 notions (and their antitheses) of artistic style: 1. linear-painterly, 2. place-recession, 3. closed-open form, 4. multiplicity-unity, 5. absolute-relative clarity

and connections between artists. The paper fine-tunes learned features on the task of painting style classification. Such features are used to assess the similarity to other paintings based on a ground truth of temporal sequence. They discover that learned features perform best at the task and present some new, plausible, correlations. Castellano and Vessio, 2021 reconstructed influences between authors based on pre-trained VGG16 features of the artworks and a nearest neighbour search for the most similar artworks to each anchor artwork. This work, although rather basic methodologically, constitutes a valuable operationalisation of the problem of pattern recognition.

An additional effort, that is closely related to the work in this thesis, is that by B. Seguin et al., 2016, who builds an image retrieval model to discover artworks with common visual patterns to the input image. The research, part of the Replica project[IV], proposes a triplet learning architecture (with ResNet or VGG) fine-tuned on the Cini Photo-archive and Web Gallery of art. The results of the paper are remarkable for the time, with a mean average precision (MaP) of 76.6 achieved in Seguin's PhD dissertation in 2018 (B. Seguin, 2018). In addition, Shen et al., 2019, in the effort of finding near duplicate patterns in art, focuses on a self-supervised learning based on spatial consistency in the Bruegel's dataset. They discover a considerable number of near duplicate patterns from the LTLL and Oxford5K datasets. An overview of this research context can be found in 1.1.

Interestingly, we notice that all the efforts surveyed, even the most recent ones, utilise ResNet, VGG or InceptionV3 in their experiments, without introducing more recently released models with lower error levels on the object recognition tasks of the Large Scale Visual Recognition Challenge (ILSVRC, Krizhevsky et al., 2012).

1.4 Thesis contribution

Tracing visual patterns in art history introduces some difficulties that go beyond those of image recognition tasks and even artist, style, and genre detection. A group of artworks sharing the same pattern may include some preparatory drawings and their transformation into the finalised artworks (a sketch becoming a fresco, a sanguine drawing into a painting); the realisations may be accomplished on different media (a statue and a painting); the propagation can take place across centuries, bearing traces of the diachronic evolution in the figurative style; or across Europe, for instance, translating the German compositions into the Italian modern style; the patterns may have evolved outside of their native iconology, with a Cleopatra transforming into a Lucretia (B. Seguin, 2018).

While a model that detects pattern propagation has to be aware of style, genre, material and iconography, it must also learn to be invariant to such changes. The model must be resilient to ample scale changes (when a section of one work becomes the leading figure in another), to mirroring (which can often be found in etchings), to dusting and deterioration through time

[IV] The Replica project will be explained and referred to throughout the thesis

and to many additional attributes.

In addition to the difficulties related to the task of pattern recognition mentioned above, the contribution of this thesis lies in the semi-supervised clustering approach.

While image retrieval does not manage an upper boundary of what is to be considered a pattern, when clustering, i.e. extracting the groups of artworks that share the same pattern, the model must be able to determine the border between pattern and non-pattern without human supervision. This is particularly difficult if we consider the range of variations that exist in copies and those that belong to different acquisitions of the same physical object. Think of conservation periods, additions, changes in perspective of acquisition, lighting, particulars.The model should not include different acquisitions of the same material object as in an ideal world, however, this is in practise not the case and deduplication methods on the task are not trivial. Therefore, the clusters can be misled by the presence of such undetected duplicates.

Additionally, some iconographical areas of art history are very densely populated as, for example, the portrayal of the Madonnas. The invention space in these areas is limited in comparison to its density and a vast number of configurations have been largely explored. The clustering should therefore be able to distinguish when two such artworks are distant enough to not form a pattern any more. An even more complex case is that of the crucifixion, where the variance in the body posture is extremely limited by physical reasons.

The complexity of the task is even more considerable if we assume that we have no preemptive knowledge of the amount of patterns in the dataset, and on what portion of the dataset belongs to any pattern group. We know that the clustering should not be limited to exact copies, but where should it stop?

While previous research focused on image retrieval tasks or graph representations of author influences, this thesis attempts to identify clusters of patterns. The shift from image retrieval to clustering allows to immediately detect the artworks that share a pattern, without having to rely on an input image search. With respect to artwork and artist influence graphs, of which clustering is an essential step, we provide a task-tailored approach to producing the clusters, unlike previous attempts, which are, on the contrary, based on generic pre-trained networks. The task-tailored clusters produced should thus reflect specifically groups of patterns and can likewise be used to produce more accurate artist influence graphs.

Tracing the evolution of such pattern groups sheds light on innumerable questions of fundamental relevance for art history. These include delineating the origin of the pattern, recognising artistic figures and workshop dynamics, assessing the authorship of each piece, recreating a network of connections, following the provenance of the pieces. More broadly, it allows to investigate the border between copies, forgeries, imitations and serves to better comprehend the cultural and socio-economic sphere of the time.

In such endeavour, the thesis assesses the feasibility for a machine to automatically suggest groups of artworks that share the same pattern, thereafter referred to as *pattern groups*. Furthermore, it shows its potential through an iterative annotation process and, ultimately, represents this continuum of modifications of patterns in a 2D space using automatically extracted coordinates.

In this thesis, we begin with an introduction to the Cini Photo-archive and the annotated set of patterns used by B. Seguin et al., 2016. We present the methods adopted, subdivided in feature learning (inspired by the work by B. Seguin, 2018), feature clustering and retraining. In the following section, we present the interface used for annotation and visualisation followed by a quantitative and qualitative analysis of the results and findings. We conclude the thesis with a discussion on this work and its process.

2 Dataset

To trace the propagation of patterns in art history, one cannot prescind from the nature and amount of artworks available for use. Recent years have witnessed an unprecedented availability of photographed art, with datasets from the Alani et al., 2018, Wikiart, Tate, MoMA, and many more appearing on the web.

As part of the Replica project, a groundtruth of patterns was compiled, which includes predominantly artworks from the Cini Photo-archive and the Web Gallery of Art. Such a groundtruth, to the extent of our knowledge, is the richest in pattern propagation. For this reason, we adopt this groundtruth in the learning. It is crucial for this thesis to adopt a vast dataset of artworks with a similar distribution to that of this groundtruth (B. Seguin et al., 2016). Therefore, we opted for the Cini Photo-archive for our analyses. In this respect, this section briefly presents the data, the groundtruth and the pre-processing steps.

2.1 The Cini Photo-archive

For the most part, this projects employs the data of the Cini Photo-archive itself. The photo-archive (*fototeca*) contains over 730'000 photographs, including the widest testimonial of Venetian art, as well as art from other Italian and European regions (B. Seguin, 2018).

The Foundation has digitised roughly half of the collection as part of the Replica project. The digitisation took place between July 2016 and August 2017 and during which time the foundation digitised all the photos that were collated in large cardboards (*schedoni*). These cardboards contain information on the object of the photograph (the author and description of the photographed piece), and details about the photographer, place and time the photograph was taken.

B. Seguin, 2018 extracted a number of metadata fields from each digitised cardboard. The metadatum on the author was matched with existing artist names databases (Wikiart and Union List of Artist Names), which allowed to incorporate more metadata fields about the authors to the available data.

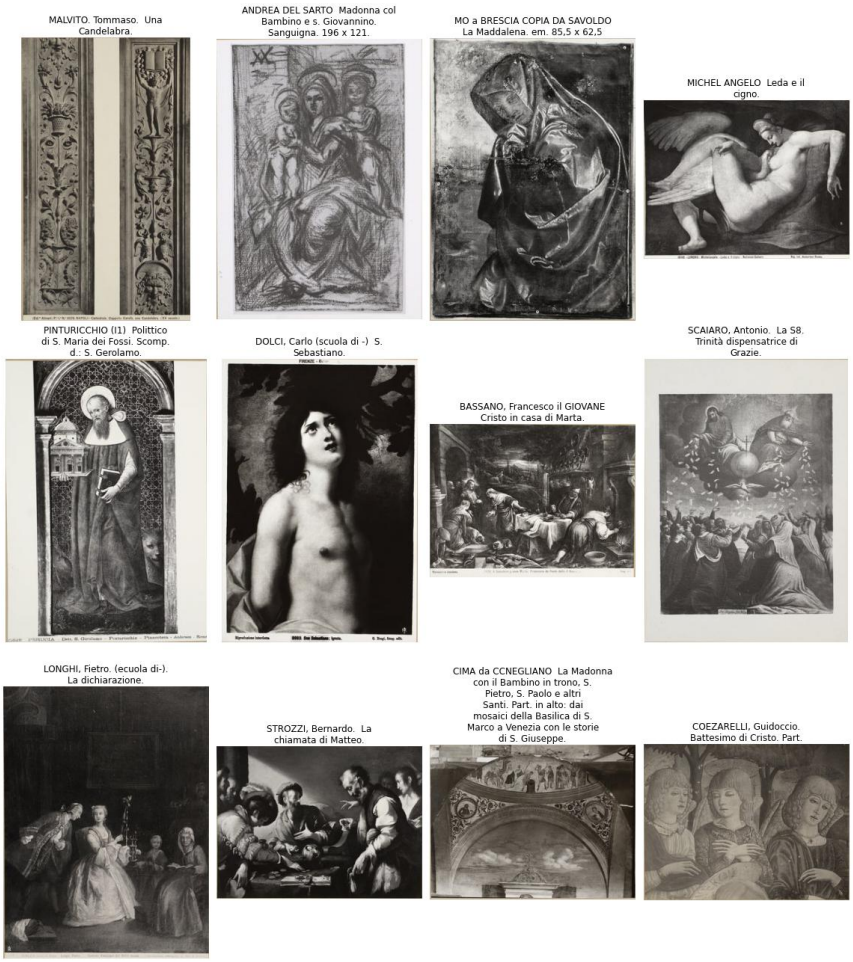


Figure 2.1: Sample of artworks in the Cini photo-archive. The artworks are in black and white or sepia. We find sketches, paintings, frescoes, but even door details.

The final collection contains artworks spanning from the XIII to the XX century, with the bulk of the data residing between the XIV and the XVIII century. The artists in the collection are about 50'000, with the most represented artists in the collection being, in order, Tiepolo, Guardi, Tintoretto, Michelangelo, Veronese, Tiziano, Raffaello, Palladio, Giotto and Bellini.

We observe that 74% of the artworks are Italian, with French, Flemish / Dutch, English and German art representing each more than 2% of the archive. The photographs are mainly in black and white. These are taken from different perspectives, including some close ups on details, some photos of frescoes taken from a distance, some images with the frame and some without. The dataset contains also numerous architectural photographs of buildings, interiors and design elements. A section of the photographs even include photos of book pages. Furthermore, the collection contains a large number of duplicated artworks[I], especially Guardi's, Tiepolo's, Tiziano's and Bellini's works. Numerous are also the pattern groups, which

[I] Multiple photos of the same physical artwork or serial print

Figure 2.2: Example of a cardboard. Retrieved from the Cini website (top) and Replica, 2018 (bottom). Each cardboard contains a photo and about six fields with metadata on the top of the *schedone*.

will be elucidated in the following section. A table with few relevant statistics can be found in Table 2.1.

2.2 The groundtruth

2.2.1 The formulation

The groundtruth dataset produced for the Replica project, which we will refer to as the *morphograph*, is of greatest value for this project. The morphograph is a graph-like structure that encodes the relations of similarity between artworks (as in Figure 2.3). In fact, the term morphograph indicates the graph representation of such annotations on the collection. These annotations can be *positive* or *duplicate*. A duplicate relation between two photographs indi-

Author	Author Nationality	Author Period	Description	Author with most **duplicates**	Author with most **patterns**
Tiepolo	Italian	XVI century	Facciata	Guardi	Tiziano
Guardi	Dutch	XV century	Veduta	Tiepolo	Cranach
Tintoretto	French	XVIII	Esterni/Interni	Tiziano	Bassano
Michelangelo	Flemish	XVII century	Madonna col Bambino	Bellini	Raffaello
Veronese	German	XIV century	Autoritratto	Raffaello	Da Vinci

Table 2.1: Table with summary statistics of the collection. The table presents the most common artists, nationalities, periods, artwork content, artists of which the collection features most duplicates and most patterns. The fields are in descending order.

Figure 2.3: Illustration of the graph nature of the morphograph. We observe how the images that share a pattern are connected with each other by an edge, while those that are not do not share any edge.

cates that two images are two distinct photos of the same physical object. A positive relation, on the other side, indicates that the artworks in the two photographs share a *visual pattern*.

This project will focus on the first type of relations, the positive ones. In this setting, we formalise the morphograph as:

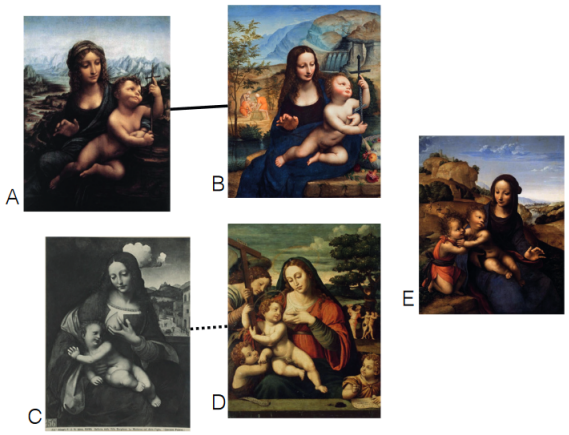
$$\mathcal{G} = \{(A, B) \in \mathcal{C} \mid A \sim B\} \tag{2.1}$$

where $\mathcal{C}$ is the collection of photographs and $\sim$ represents the similarity between A and B. In this case, we consider two artworks similar if they share the same pattern. The photographs in the collection that belong to the morphograph are the *nodes* of the graph and they share an *edge*, a connection, in the graph if they propagate the same pattern. We will adopt this terminology throughout this thesis. Formally, we constrain the similarity in the graph as follows:

The equation 2.2 on the *partial ordering* of the graph formulates that, given an image B that

$$\forall B \in \mathscr{C}, \forall C \in \mathscr{C}, ((A\text{–}B) \in \mathscr{G} \& (A\text{–}C) \notin \mathscr{G}) \Rightarrow (A\text{–}B) > (A\text{–}C) \tag{2.2}$$

Figure 2.4: Equation from B. Seguin, 2018



c_1	c_2	$c_1 <_1 c_2$	$c_1 <_2 c_2$	$c_1 < c_2$ (with $<$ representing $<_1 \cap <_2$)
$(A\text{–}B)$	$(B\text{–}E)$	$>_1$	$>_2$	$>$
$(D\text{–}E)$	$(C\text{–}D)$	$<_1$	$<_2$	$<$
$(A\text{–}B)$	$(C\text{–}D)$	$>_1$	$?_2$	?
$(B\text{–}E)$	$(D\text{–}E)$	$>_1$	$<_2$	?

Figure 2.5: Partial ordering of the morphograph. If A is connected to B and C to D, we observe that A-B is closer than A-E and D-E is more distant than C-D. Retrieved from B. Seguin, 2018

shares an edge in the graph with A and the image C that does not share an edge with A, we can conclude that A and B have a stronger similarity than A and C, or, in other words, that A and B share the same pattern, while A and C do not. We will come back to this formulation when creating the dataset for learning.

We indicate as $N(A)$ the *neighbourhood* of the node A, defined as the set of nodes that have a direct edge with A. Furthermore, we define a *connected component* in $\mathscr{G}$ as the complete set of nodes such that from any A and B in the connected component there is always a path between A and B. This terminology will be adopted in the Methodology and Results sections.

2.2.2 About this morphograph

Considering the morphograph with only the positive connections, we count 7284 connections between artworks (edges of the graph) and a total of 2175 unique artworks (nodes of the graph). Unfortunately, only 4900 of the connections contained artworks that could be accessed to date, counting to 1800 available works.

In addition to the digitised data from the Cini Foundation, the project Replica adopted the

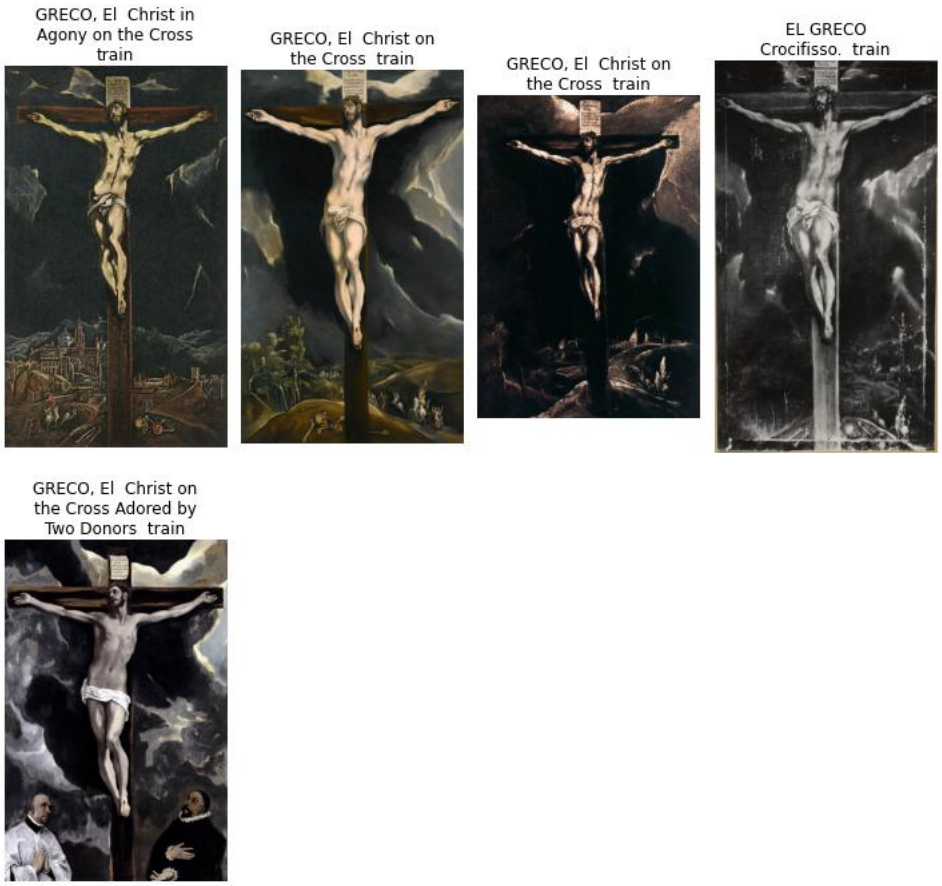


Figure 2.6: Example of a pattern group with autograph variations. El Greco, Christ on the Cross group.

Web Gallery of Art (WGA) database in making the morphograph. The database began in 1996 as a site for Renaissance art and is now a virtual museum of European fine arts, containing over 52.800 copyright free artworks (from WGA website). Of the morphograph, 1322 images are from the Cini Foundation and the remaining ones are from the Web Gallery of Art.

The connections were annotated by a team of art historians during the Replica project. An annotation platform was created during the project to facilitate the process of creating the morphograph. Originally, a small number of interns of the project, art history university students, conducted a thorough research on the known patterns of Venetian art and annotated them using the platform. Successively, the remaining annotations were inserted based on the suggestions by the platform itself (B. L. A. Seguin, 2018). This process will be explained more in depth in the Interface section.

The resulting morphograph contains 418 pattern groups in a wide variety of forms. The groups span from etchings, engravings, drawings, to paintings of architectural vedute, to autograph

Figure 2.7: Example of a pattern group with vedute. Canaletto, Veduta of Palazzo Ducale and Piazza San Marco group.

variations on a work. Few groups include preparatory drawings and their successive painting by the same author, as well as copies from the bottega, the scuola or inspiration from other authors.

Let us look more in depth into a few examples to better comprehend the nature of the morphograph. The first pattern group we will consider is the epitomical half-laying woman (Figure 2.10). This leitmotif, whilst originating already in the ancient times, has gathered attention during the Venice of the XVI century. The pattern originated from Giorgione's Venus, a warm coloured, sleeping Venus, with the elbow over her head. The figure is inscribed in a natural environment, and is unaware of the human gaze. The same motif appears in a number of copies from the period and drawings by Valentin Le Fevre (as in Figure 2.10). A few decades later, Tiziano reworks the composition with the master, depicting the awakening of the Venus, who is now looking attentively and slightly provocatively at the viewer from the comfort of her room (Kultermann, 1990). Other pieces by Tiziano introduce Cupid in the scene, who is later given a rather central role in the compositions by Bourdone. Also Cranach the younger, as Tiziano, gradually makes up the nymph depicted by the master Cranach the Elder. Thereafter, the representation of the Venus from Giorgione's seed branches out, reaching a variety of forms and iconographies.

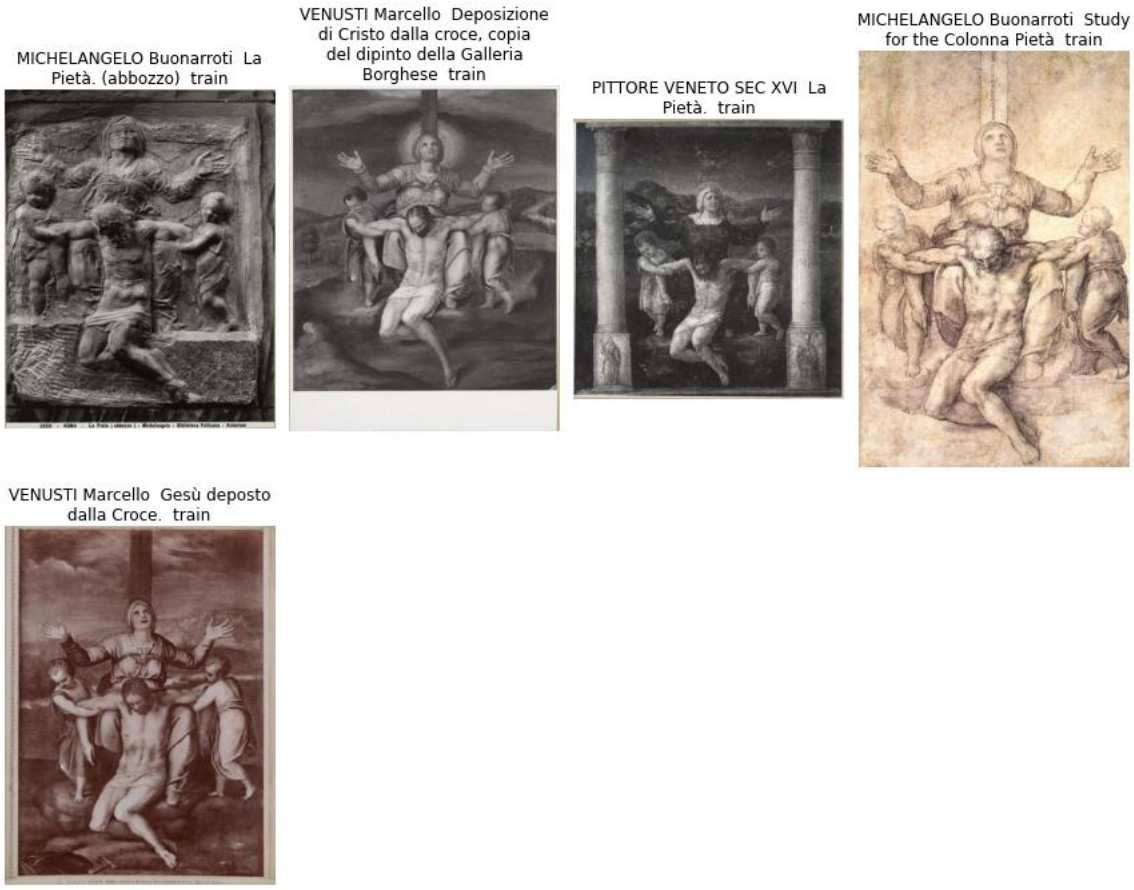


Figure 2.8: Example of a pattern group with preparatory drawings and different materials. Michelangelo's Pietà group, with painting by Venusti.

Juxtaposing to the great variety of the Venus group, El Greco offers a number of groups of containing solely variations of his autograph work. An example is the crucifixion in Figure 2.6. The compositions are catalogued by the Cini as autograph variations of the same pattern, featuring different colour compositions and landscapes and even figures. El Greco's oeuvre is filled with such variations, but also copies in bottega and outside (Pozzolo, n.d.). This, together with the notoriety that the painter gained in the first decades of the XXth century, contributes to a large contamination of copies, forgeries but also patterns of his works(Pozzolo, n.d.). Unfortunately, the collection used in this thesis covers a period that predates the patterns of the XXth century and we are, therefore, unable to trace the full evolution of such patterns.

The contribute to the complexity and variety of the morphograph, we observe many examples of works of the Vedutismo period in Venice. We see the example of Canaletto's series of paintings of the Grand Veduta of Piazza San Marco in Figure 2.7. The architectural veduta, in these works, is the undisputed centre of the depiction. The veduta is drawn from the other

side of the canal, featuring the ever changing activity of the gondole and boats in the water. The canal is represented from slightly different perspectives. This last attribute of changing perspectives inside the same group appears in the morphograph almost solely for architectural vedute, whilst the rest of the works generally share the same perspective.

Another frequent element in the morphograph is the presence of copies that were produced by the scuola or bottega of the original author. One of the prolific cases of such phenomenon can be found in the scuola bassanesca (Ericani et al., 2010) of the XVI-XVII century. In Figure 2.9, Jacopo Bassano proposes a scene of the 'Adorazione dei Pastori', a pastoral rural scene which is typical of his style. In the pattern group, we see copies by other family members, the son Leandro Bassano, as well as other members of the scuola.

Finally, an important example that manifests the difficulty in the tracing such patterns, is presented in Figure 2.8 (Capelli, n.d.). The group features different materials, in this case, canvas, yellowed sheet and stone/metal (?), and, consequently, different techniques (engraving, sketch and painting). The group in Figure 2.8 is only a section of the over 15 known copies of the work. The pattern depicts the desperation of the Madonna upon the removal of Christ from the cross, who is now held from falling by two angels. The pattern in Figure 2.8 features what is probably the original drawing by Michelangelo and his successive bas-relief. The paintings, on the other side, were inspired by the drawing and translated to painting by Venusti. This dynamic is not uncommon in Michelangelo's work, who often trusted the interpretation of Venusti in the translation to painting. The group is still an open subject of attributions (Capelli, n.d.).

2.3 The subset

Given the massive size of the Cini Foundation and WGA database (370106 images) and the limited time and compute power, only a subset of the data was used for training the models for this thesis. In creating the subset, we paid attention to maintaining a distribution similar to that of the whole collection whilst retaining only the relevant artworks.

In creating the subset, the following steps were undertaken:

- Only artworks produced (roughly) between the XV and the XVIII century were considered, as these are the most represented in the morphograph. The artworks outside this period were excluded as outliers due to the very scarce representation in the morphograph. This lead to a set of 155004 images.
- The data was deduplicated based on the 'duplicate' link in the morphograph. For each group of duplicated artworks only the first one was kept. Deduplicating helped reduce the overhead in the collection whilst maintain the entire information signal. We ended up with 122901 images after this step.
- The data was again deduplicated by maintaining only the first photograph of the groups

of artworks with the exact same *Author*, *BeginDate* and *Description* fields. This step potentially eliminates non-exact duplicates but it was considered necessary after the first results were found to feature a large percentage of duplicates (or exact copies). We obtained a set of 111563 images.

- The photographs whose *Description* included

```
'tavolo',
'facciata',
'paesaggio',
'esterno',
'pag.',
'soffitto',
'sedia',
'veduta',
'scorcio',
'capitell',
'architettonic',
'fontana',
'edifici',
'libr',
'architettonic'
```

were eliminated. In fact, we are not interested in landscapes, facades and book pages or any architectural element in this thesis. This step yielded a set of 101894 remaining images.

- The photographs whose cardboards are physically stored in the folders from 70A to 86C were also excluded as these folders contain architectural photographs. This resulted in a cleaned set of 78078 photographs from the Cini (95671 total).
- 10000 of the cleaned set from the Cini were sampled at random the create the subset.
- The images in the morphograph were added to the subset (including WGA), resulting in a cleaned sampled set of 12586 photographs.
- The images in the cleaned set that were no longer available were removed, resulting in **8959** photographs.

From this moment, we will refer to sampled set of 8959 photographs as the collection $\mathscr{C}$. This set will be used for training, validation and testing as explained in the Methodology section. Finally, we adopt the larger set of 78078 images for prediction for the final clustering as explained in the Results.

Figure 2.9: Example of a pattern group including original, copies and works from the scuola. Jacopo Bassano's Adorazione dei Pastori (and scuola).

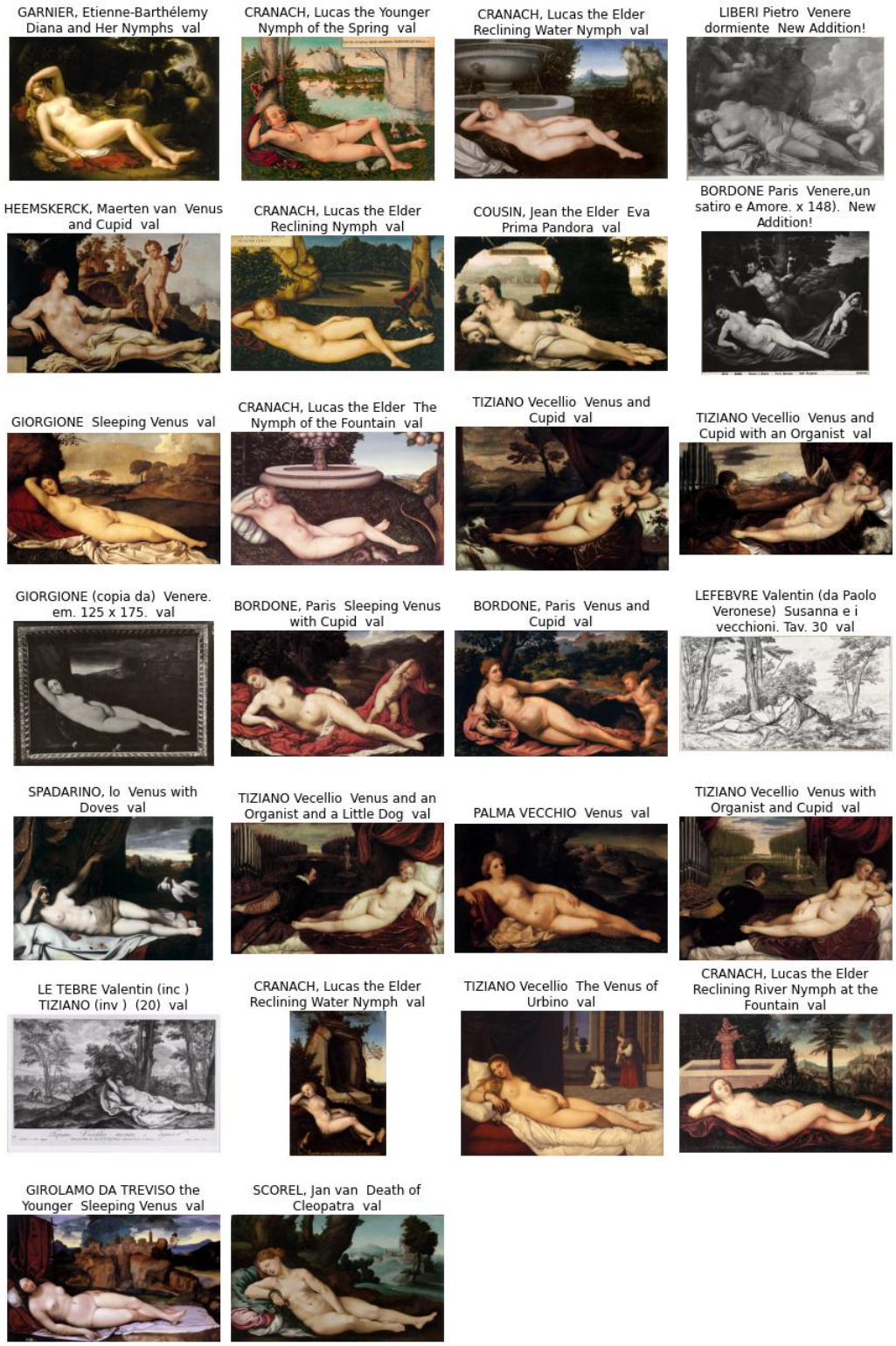


Figure 2.10: Example of a renowned pattern group with many authors. Giorgione's Sleeping Venus group, featuring Tiziano, Bourdone, Le Fevre, Cranach.

3 Methodology

As previously discussed, this thesis heavily relies on the efforts by B. Seguin, 2018 to derive visual signatures of artworks for image retrieval. His pipeline is reworked in this thesis and used to ultimately produce clusters of the visual signatures.

In this section, we first present a review of papers that discuss image retrieval networks. Successively, we describe the model by B. Seguin, 2018 and our changes. We outline previous efforts that take advantage of visual signatures for image clustering and dive into the clustering models adopted. Finally, a small section is dedicated to the semi-supervised approach to learning pattern clusters.

3.1 Convolutional Neural Networks: a bare minimum

Convolutional neural networks (CNN's) were introduced by LeCun in 1989 and earned a great deal of success in the field of computer vision and beyond (Goodfellow et al., 2016). These networks are apt for processing 2D grid-like information (such as images) using on the convolution operation, which gives it the name.

When using CNN's, standard practises involve selecting the architecture, the transfer learning, the pooling, the loss; but also choosing the activations, the optimiser, scheduler, regularisation, the data augmentation, the initialisation, the dropout and much more (Goodfellow et al., 2016). In particular, in this thesis, we focus on the selection of the adequate architecture, the transfer learning, the loss and the data augmentation.

The choice of architecture is the selection of the sequence of units, activations and connections between layers that comprise the final network. A common approach to architecture selection consists of adopting already existing networks (such as ResNet, VGG and InceptionV3 which we already mentioned) and initialise them using learned weights. These architectures and their weights are commonly referred to pre-trained architectures.

The pre-trained architectures are fine-tuned using transfer learning. Transfer learning (TL)

is a deep learning technique used, predominantly, to improve the performance of a model when the training data is scarce and to generalise the model to a specific task[I]. It consists of continuing the training of the pre-trained architecture on the training set selected for the task.

Another method to improve the performance of a model when training data is scarce is data augmentation. This consists of selecting a set of steps of manipulations of the input image to induce learned invariances in the model. These include generally random cropping of the image, random colour jittering, flipping, rotating and normalising (Goodfellow et al., 2016).

Finally, the choice of the loss is essential in a deep learning task. In fact, the loss determines in which cases the model learns from the training images and how. Common choices of loss functions are the squared loss, the cross-entropy loss, the Hinge loss. The choice depends essentially on the nature of the task (i.e. regression, classification).

3.2 Triplet learning for image retrieval

In this thesis, we initially learn and assess the model on the task of image retrieval. Image retrieval (IR) with CNNs has been largely studied by the literature (Babenko et al., 2014; Balntas et al., 2016; Gordo et al., 2017; Ho et al., 2021; Jahrer et al., 2008; Radenović et al., 2018; Simo-Serra et al., 2015; Tolias et al., 2016). It is a particular case of information retrieval that finds relevant images in a collection $\mathscr{C}$ based on an input image I. The task of IR is formalised as:

$$r(g(I)) = r(f_I) = [\arg\min_{x \in \mathscr{C}}(d(f_I, g(x)), .., \arg\max_{x \in \mathscr{C}}(d(f_I, g(x))],$$

$$g(x) : \mathbf{R}^{r_1 \times r_2} \rightarrow \mathbf{R}^{x \times 1},$$

$$f(x) : \mathbf{R}^{x \times 1} \rightarrow [\mathbf{R}_0^{r_1 \times r_2}, \mathbf{R}_1^{r_1 \times r_2}, ..., \mathbf{R}_N^{r_1 \times r_2}]$$

The input image I (of size $r_1 \times r_2$) is mapped to a feature descriptor f_I of size x using a function $g(x)$. The descriptor is used as query to retrieve the images. The image retrieval function $r(x)$ maps the input query to an ordered list of relevant images that belong to the collection $\mathscr{C}$, where the image $x \in \mathscr{C}$ with the descriptor of lowest distance d to that of the query is retrieved first, until the most distant descriptor to f_I which is retrieved last. Commonly, the ordered list is cut to the first N results for computational reasons.

In particular, a CNN based image retrieval model maps an input image I to a fixed length descriptor f_I (of size x) of lower dimension. This is done with by a number of successive convolutional blocks that reduce the 2D image-level representation I to a 3D spatially compressed

[I]Formally speaking, in a supervised learning setting with an input target space X, an output space y and a function $F : X \rightarrow y$ that minimises the expectation of a given loss, TL adds a second input space X', the source space, and a second output space y'. Normally, at least one of $X \neq X'$ and $y \neq y'$ is true. The goal of TL is to find a better F' that exploits the source data and, possibly, the target data (Goodfellow et al., 2016)

representation F, that is globally pooled[II] into a 1D descriptor f_I as:

$$g(I) : \mathbf{R}^{r_1 \times r_2} \rightarrow \mathbf{R}^{d_1 \times d_2 \times d_3} \rightarrow \mathbf{R}^{x \times 1}$$

where ordinarily $r_1 \times r_2 << d_1 \times d_2$ and $d_1 \times d_2 \times d_3$ are the dimensions of F, selected before the training.

In what follows, note that $g(x)$ is the CNN model, and f_I the (compact) image descriptor.

3.2.1 The literature

To obtain the compact image representation needed for image retrieval, previous work has focused on large visual codebooks (Bag of Words), Fischer and SIFT vectors and other methods (B. Seguin et al., 2016). Recently, the field has shifted towards the adoption of CNN models as mapping function $g(x)$.

The use of CNN models for image retrieval consists of some essential steps that will be reviewed in this section. These are: *feature descriptors learning, global pooling, hard negatives sampling,* and *whitening*. This last step is optional and consists of reducing the dimensionality of the feature descriptors to improve and optimise the retrieval.

In light of the great success of transfer learning, feature descriptors are generally obtained by fine-tuning models pre-trained on ImageNet (Krizhevsky et al., 2012). Babenko et al., 2014 investigates which layer of the pre-trained model should be used as feature descriptor, indicating the last convolutional layers after the global pooling as most apt to the task.

The fine-tuning is carried out either through a pair based or triplet based learning (Jahrer et al., 2008; B. Seguin, 2018; Simo-Serra et al., 2015).The pair based model was introduced by Jahrer et al., 2008. It uses a siamese architecture to learn, with contrastive Hinge loss, to reduce the distance between pairs of samples that share a positive connection (label 1) and to increase the distance between pairs that with a negative link (label -1). The triplet model was created as an extension of the siamese network to learn concurrently the positive and negative samples. In fact, the triplet model receives as input an anchor image, a positive image and a negative one; the positive image is an image that shares a positive link with the anchor, while the negative shares a negative link with the anchor. The loss of the triplet model is the triplet loss, which is a *ranking Hinge loss with margin*.

A common issue with triplet models is that negative samples are often not causing any learning, as they are often already clearly more distant than the positive samples to the anchor. To solve this issue, Simo-Serra et al., 2015 introduced the *hard negative sampling* technique, which samples negative images among the set of images that are very similar to the anchor, but are not positive samples. The procedure was refined by Gordo et al., 2017's brute force approach,

[II] Global pooling is the operation that compresses the 3D signal into a 1D signal using aggregation operations such as maximum or mean

which sampled all possible triplets and sub-selected only the ones that would incur in a loss for training.

With the same objective, Balntas et al., 2016 proposed the *anchor swap* technique which swaps the positive sample with the anchor whenever the distance of the positive with the negative sample is smaller than that of the anchor with the negative sample. This allows the model to always learn from the most difficult configuration of the three inputs. Further variations to the loss were proposed by Ho et al., 2021 who introduced a *double margin loss* that regulates, together with the inter-sample distance, the intra-sample distance between the anchor and the positive sample.

Some researchers proposed complex activations[III] for the last layer of the network, which is used for the compact representation of the image (Gordo et al., 2017; Tolias et al., 2016). Tolias et al., 2016 proposed the regional maximum activations of convolutions (*R-MAC*), for which the model max pools over different regions and sum pools over the various regional descriptors to obtain a region aware compact representation of the image. This was end-to-end optimised by Gordo et al., 2017 with a system for proposing Regions of Interest (RoIs) during training.

Finally, most works whiten the results using Principal Component Analysis (PCA) (Radenović et al., 2018), or even include the whitening in the training (Gordo et al., 2017).

3.2.2 Replicating Replica with some key changes

B. Seguin, 2018 is, to the best of our knowledge, the first attempt at image retrieval for pattern propagation in art. Given the remarkable results obtained at the task, this project begins with the re-implementation of the model in B. Seguin, 2018 from `Tensorflow 1.x` to `Pytorch 1.11`, and proceeds with extending the project with new methods and analyses.

The base architecture

Among the top competitors of ILSVRC, this paper uses experiments with different architectures and their pre-trained weights as starting points for the retrieval model: ResNet, ResNeXt, and EfficientNet (K. He et al., 2016; Tan and Le, 2020; Xie et al., 2017). The saved architectures and their weights are downloaded and imported using the `torchvision` module and the models are then fine-tuned on the morphograph.

From the graph in Figure 3.1 we observe that computer vision models have largely improved in object recognition tasks since the advent of ResNet-101. The models proposed after 2016 consist largely of clever variations of the ResNet, including ResNeXt, or of EfficientNet based models or scaling factors. Some architecture, as of 2020, exploit the first transformer methods for images. In these experiments, we focus on CNN based methods as transformers have

[III] Activations are functions, present at each layer of the network, that introduce non linearity for the learning. Common choices of activation are ReLU and sigmoid.

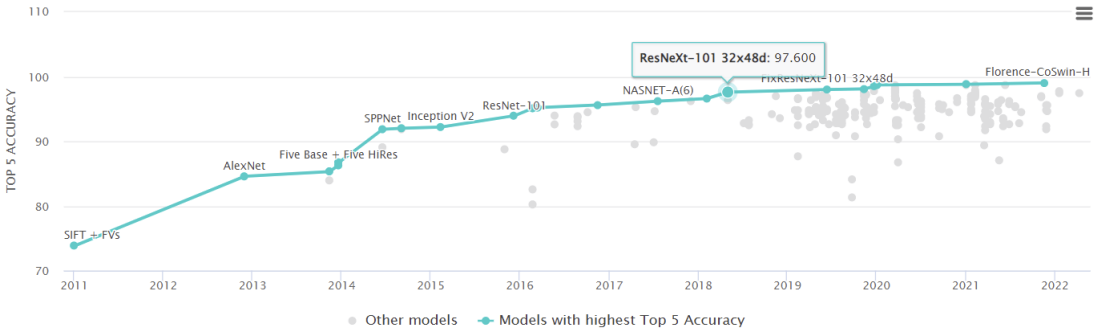


Figure 3.1: Top competitors for ILSVRC (until 2022). Retrieved from paperswithcode.com. The figure shows the performance of the top competitors of the object recognition task of the ILSVRC. We see ResNet in 2016 achieving the highest top 5 accuracy for the time. On its right, we observe an increase in performance by models such as NASNET, ResNeXt and more.

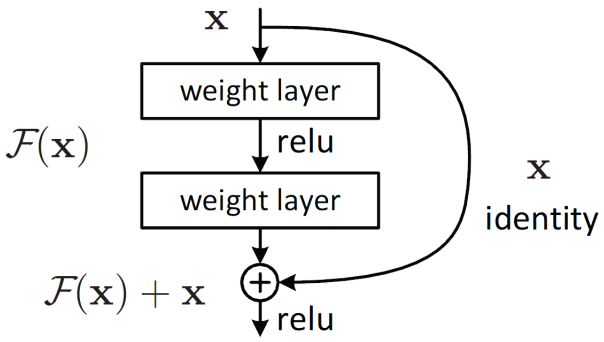
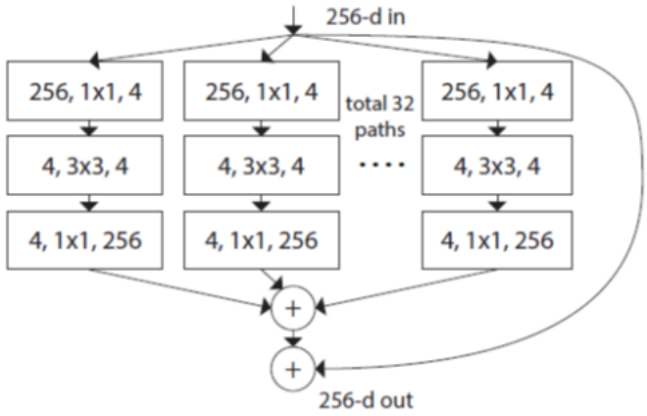


Figure 3.2: ResNet residual block. Retrieved from T. He et al., 2018. The figure visualises the skip connections with identity mapping which have been proposed in T. He et al., 2018

prohibitively large architectures for our resources (reaching 10000 million parameters). In what follows, we attempt to quickly overview the three methods.

ResNet is an extremely deep cascading network that learns via micro residual modules. In particular, each *residual module* learns, subsequently, a 1x1 bottleneck, a 3x3 and another 1x1 bottleneck residual and it adds it to a skip connection (an identity mapping). The residual block of ResNet can be seen in Figure 3.2. The use of residuals and identity mappings is introduced to avoid the vanishing gradient problem and to allow the architecture to adapt to the complexity of the task or data. This is achieved by learning to use only some portions of its representational capacity (K. He et al., 2016).

ResNeXt blocks also act as a residual block. The model consists similarly of consecutive convolutional blocks with each block sharing the same hyperparameters, and reducing the spatial dimension by a factor of two, while increasing by two its depth. Differently from other architectures, ResNeXt introduces a *cardinality* level in each block, next to the width and the depth. The cardinality is to be interpreted as the number of complex transformations that are



$$\mathcal{F}(\mathbf{x}) = \sum_{i=1}^{C} \mathcal{T}_i(\mathbf{x})$$

Figure 3.3: ResNeXt residual block. Retrieved from Xie et al., 2017. Differently from ResNet, we observe how each block passes through 32 paths (C=32 is the cardinality) concurrently.

applied for each convolution (the cardinality can be seen in Figure 3.3). This was proven to improve the model performance more significantly than any increase in the width and depth (Xie et al., 2017).

Finally, *EfficientNet* is a family of models introduced by Tan and Le, 2020 that was created by optimising convolutional architectures for accuracy and FLOPS (FLoating point Operations Per Second). Particularly, Tan and Le, 2020 formalised the *scaling* operation on neural architectures. They introduce a uniform scaling to the width, depth and resolution of the models based on a compound factor. The architecture and scaling procedure can be seen in Figure 3.4.

Based on the results by Babenko et al., 2014; B. Seguin et al., 2016, we use the architectures mentioned above until their last convolutional layer. We include a mean global pooling layer as:

$$f_{mean}(I)[l] = \sum_{j,k} F_{j,k,l}$$

and normalise the descriptor with L2 normalisation as:

$$f_I = f_{norm}(I) = \frac{f_{mean}(I)}{||f_{mean}(I)||^2}$$

The normalisation creates a descriptor that is suited for the similarity computation of image retrieval.

The procedure is shown in Figure 5.24.

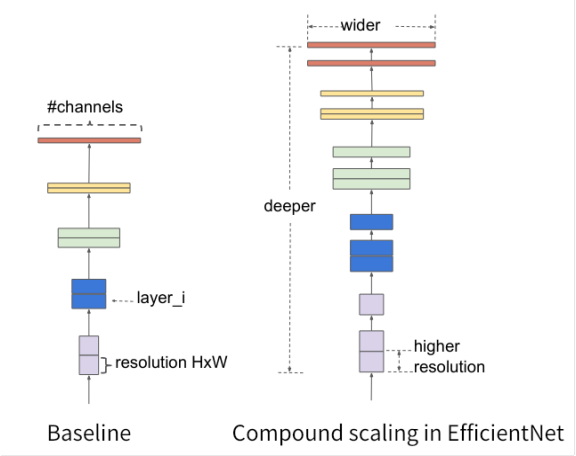


Figure 3.4: Efficient base architecture and scaling. Retrieved from Tan and Le, 2020. The figure shows EfficientNet's architecture, including a visual representation of the automatic scaling with compound factor that was introduced by Tan and Le, 2020.

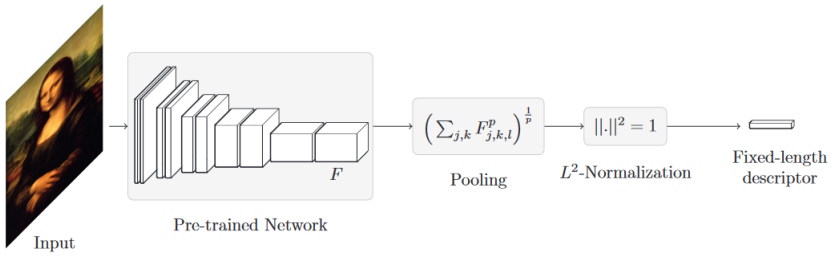


Figure 3.5: Replica model architecture. Retrieved from B. Seguin, 2018. From left to right, the 2D image is passed through a CNN model that maps the descriptor to the 3D compact representation F, which is pooled to a 1D descriptor and normalised to obtain the final fixed length descriptor used as query.

The triplet model and loss

As mentioned, a triplet model receives as input three images, an anchor a_i, an image similar to a_i, b_i and an image different from a_i, c_i. The input triplet is commonly referred to as the *anchor, positive and negative* triplet, a_i, b_i and c_i respectively. The triplet model is trained using a Hinge margin loss that conceptually pushes image a_i closer to b_i and farther away from c_i when this order is not already respected.

As one can see in Figure 3.6, the three images are each mapped to a descriptor f_I. For simplicity in the notation, we refer to f_{a_i} as A_i, f_{b_i} as B_i, and f_{c_i} as C_i. The three descriptors are fed into the Hinge based margin loss. The loss computes the *positive distance* $d^+ = d(A_i, B_i)$ between A_i and B_i and the *negative distance* $d^- = d(A_i, C_i)$ between A_i and C_i. For each triplet, the loss is typically calculated as:

$$\mathscr{L}_i = \max(0, m + d^+ - d^-)$$

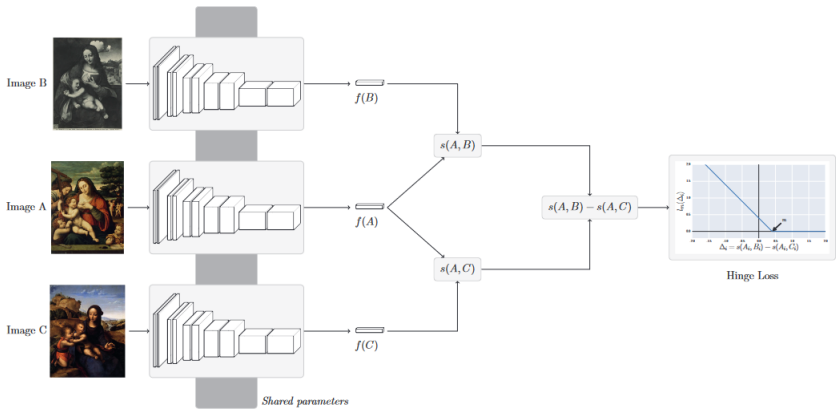


Figure 3.6: Replica triplet model architecture. Retrieved from B. Seguin, 2018. In the figure we observe the three input channels being fed into three CNN models with shared parameters (that are updated concurrently). On the right, the three output descriptors, here called f being fed into the Hinge margin loss.

This penalises the model when the anchor and the positive sample are more distant by at least a margin m than the anchor and the negative sample.

The loss adopted in this thesis incorporates to the standard loss the anchor swap technique and intra-sample penalisation (Balntas et al., 2016; Ho et al., 2021). Specifically, the *anchor swap* defines the negative distance as $d^- = min(d(A_i, C_i), d(B_i, C_i))$. It considers the distance to the negative sample as the lowest distance between A_i, C_i and B_i, C_i, thus always taking in consideration the negative distance that yields the greatest learning potential. On the other side, the *intra-sample penalisation* adds a second margin loss to the training. This minimises the positive distance to be below a margin m_2 as $\max(0, m_2 d^+)$.

Including the two additions to the Hinge loss, final loss becomes:

$$\mathscr{L} = \sum_i max(0, m + d^+ - d^- + \max(0, d^+ - m_2)) =$$

$$= \sum_i max(0, m + d(A_i, B_i) - \min(d(A_i, C_i), d(B_i, C_i)) + \max(0, d(A_i, B_i) - m_2))$$

The final loss penalises the model when the partial ordering of A_i with B_i and C_i is incorrect and concurrently pushes A_i close to B_i. In particular, it ensures that A_i is closer to B_i than C_i by at least a margin m or that B_i is closer to A_i than C_i by that margin (in case the inputs are swapped). The margin m can be interpreted as the inter-sample distance between the elements that should appear together (the As and Bs) and those that do not belong to them (the Cs). Additionally, the loss ensures that the maximal distance between A_i and B_i remains under m_2, thus enforcing a maximal intra-sample distance between the elements that belong together (Ho et al., 2021).

The training with hard negative sampling

In order to take full advantage of the morphograph for training, B. Seguin, 2018, after Simo-Serra et al., 2015's initial effort, introduced an iterative hard negative sampling procedure.

Algorithm 1 Iterative hard negative sampling. We show how c_i is computed as the set of first N results of the image retrieval of a_i using the model at epoch n (g_n) subtracting to this set the images in b_i.

for $n \in [0, ..., numberofepochs]$ **do**
  for $I \in \mathscr{M}$ **do**
    $a_i \leftarrow I$
    $b_i \leftarrow N(a_i)$
    $c_i \leftarrow r(g_n(a_i))[:N] \setminus b_i$ ▷ $g_n(x)$ is the CNN model configuration at epoch n
  end for
end for

At each epoch of the training, an image a_i is selected from the images annotated in the morphograph. Every a_i is trained with all its neighbours in the morphographs $b_i = N(a_i)$'s. Since the b_i's share a link with a_i, these necessarily share a pattern with a_i. Each configuration of the pair (a_i, b_i) is trained with N c_i's. These are computed after every epoch as the N most similar images to a_i based on the descriptors created by the model that epoch. Before the first epoch, the descriptors are produced with the pre-trained model and are updated every epoch with the new descriptors of the updated models. Thus, the first N results that are not connected to a_i in the morphograph become the eligible c_is. With this procedure, c_i can be any image in the collection such that $c_i \in \mathscr{C}$ is not in $N(a_i)$ but is very similar to a_i. The choice of cs is supported by the partial ordering property of the morphograph in 2.2, which ensures that any images that is not connected to a_i is a negative connection to a_i. It is to be noted that this is not true in practise as we could not annotate the full collection. This procedure, nevertheless, is proven to be empirically sound by B. Seguin, 2018.

B. Seguin, 2018 uses cosine distance to obtain the N most similar cs while we utilise Euclidean distance. This change does not create any difference in terms of ordering of the output as the feature descriptors are normalised.

The iterative hard negative sampling techniques is formalised as in 1.

Image retrieval from the learned descriptors

Once the model is trained, the descriptors predicted by the model are used for image retrieval. This is done by storing all the descriptors of all the images in the collection into a *nearest neighbour tree*. For this step as well as for the hard negative sampling, we consider *every* image in the collection and not only those that belong to the morphograph.

We use a Balltree implemented in `sklearn` for the nearest neighbour tree structure. The

nearest neighbour tree implements the unsupervised k-nearest neighbour[IV] using a search tree. This representation allows optimised search times compared to simple brute force queries. The tree allows querying with the descriptor of an image to obtain the ranked list of the N images whose descriptors are most similar to the input query ($r(f_I)[:N]$). The Balltree, in particular, was chosen as it performs best for high dimensional input. The search in the Balltree is utilised also in the hard negative sampling querying with a_i and returning the top N c_i's.

Overall, we have outlined a number of changes from the original model in B. Seguin, 2018, these have been summarised in Table 3.1 for clarity.

Modification	B. Seguin, 2018	This thesis	Reason
training set size	$\mathscr{M}$ = 7'284	$\mathscr{M}$ = 4'900	Images no longer available
collection size	$\mathscr{C}$ = 340'000	$\mathscr{C}$ = 8'900	Limited compute power
pre-trained architecture	ResNet/VGG	ResNet/ResNeXt/EfficientNet	Experiment with more recent architectures
loss	Hinge margin loss	double Hinge margin loss with anchor swap	Facilitate training and aiding future clustering step
retrieval distance function	cosine	Euclidean	No practical difference

Table 3.1: Overview table of the changes between B. Seguin, 2018 and this thesis.

[IV] For information on nearest neighbour trees refer here

3.3 From image descriptors to clustering

Image clustering is the task of creating groups of images based on a compact representation of the images (f_I). This task, since the advent of CNNs, has largely been based on learned representations of the images. Standard models for image clustering involve adopting pre-trained image descriptors and standard clustering methods, such as K-means, DBSCAN and Spectral clustering (Ankerst et al., 1999; Bishop et al., 1995; Ester et al., 1996; MacQueen, 1967). More recent methods explore avenues that range from fine-tuning the pre-trained architectures for specific tasks, to opting for an end-to-end training and clustering approach (Das et al., 2019; Ho et al., 2021; Mukherjee et al., 2019; Nina et al., 2019; Prasad et al., 2020; Song et al., 2013. Given their better applicability to our task, we will explore these avenues further in this section.

3.3.1 The literature

To learn clusters for a specialised task, the field quickly shifted towards fine-tuning models to produce task-specific image descriptors. Many attempts have opted a triplet loss approach, followed by kmeans clustering of the descriptors (Das et al., 2019; Nina et al., 2019). A step further was taken by Ho et al., 2021 who experimented with different triplet loss modifications that are more suited for image clustering. Such modifications have been used in the triplet loss learning for this thesis as discussed above.

Notably, none of the above articles surveys a variety of clustering methods, rather they often rely on a single clustering method.

Taking an even larger step, Song et al., 2013 was one of the first attempts at integrating an end-to-end learning model for image clustering, performing k-means in the latent space of Autoencoders. Deep embedded clustering, similarly, used an Autoencoder with KL-divergence loss to learn the latent feature space. Other remarkable efforts were done both with Variational Autoencoders and Generative Adversarial Models (Mukherjee et al., 2019; Prasad et al., 2020). Prasad et al., 2020 enforced a Gaussian Mixture Model for the latent space distribution.

Unfortunately, none of the end-to-end methods surveyed learn a task specific clustering, as they are all built on a reconstruction loss of the same image. Since pattern propagation requires highly specialised representation, as explained already in the Introduction, we focus on feature descriptors learned from a triplet learning network. Thus, we use the feature descriptors learned with the pipeline in the previous section and explore different clustering methods that are suitable for such descriptors.

3.3.2 The choice of standard clustering methods

In this thesis, we experiment with three clustering techniques to determine which is most suited for pattern propagation detection. We recall that the task has some specificities worth

considering when choosing the clustering. On one side, we expect that many of the artworks do not belong to any group, but that are rather isolated cases or outliers. Methods such as DBSCAN are particularly suited for this effort as they detect outliers. More specifically, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise, Ester et al., 1996) method creates clusters based on the concept of core points. A core point is a point that contains at least a minimum number of other points in its radius of size ϵ. In our case, since we wish to have at least two elements per cluster, the minimum number is 1. The algorithm detects the core samples and composes a cluster based on the core samples that are in each other's radius and adding the other points in those neighbourhoods. If a point is in the neighbourhood of no other point, this is considered an outlier.

Furthermore, we expect a largely populated area around the centre of the space (for example, where the different representations of the Madonna's would lie), in this case, basic density based methods might not be able to distinguish inside the highly populated area and end up with a giant component[V]. For this reason, methods like K-means and OPTICS would be beneficial. The first, in fact, detects the amount of clusters suggested, regardless of their density. The K-means method requires knowledge of the number of intended clusters (MacQueen, 1967). From these, it iteratively selects centroids that minimise the inertia of the cluster based on their radial Euclidean distance. K.means only detects circular cluster shapes and suffers heavily from the curse of dimensionality[VI].

The second, on the other hand, is a variation of DBSCAN that is able to identify areas of tighter densities even in already high density areas. The OPTICS (Ordering Points To Identify the Clustering Structure method) is similar to DBSCAN with a 'relaxed' ϵ, the max_ϵ (Ankerst et al., 1999). The algorithm constructs a reachability graph for every ϵ lower than the input max_ϵ. It agglomerates clusters based on variable densities. Compared to DBSCAN, it is more prone to detecting central points as outliers while less depended on the choice of the ϵ parameter and less susceptible to the creation of a giant component (Kanagala and Jaya Rama Krishnaiah, 2016).

3.3.3 Iterative learning

The clustering step is aimed at providing automatic suggestions of groups of images sharing the same pattern migration. This allows art historians to scan suggestions that are are likely to contain patterns. To simulate the impact of the suggestions and improve the model, we introduce a human-in-the-loop approach to clustering. We first-handedly annotate the suggestions to evaluate their quality and utility for art historians. Subsequently, we use the annotations to improve the model that produced the original clusters. In this way, we continuously learn from the newly discovered pattern groups. Additionally, practical observations of an annotator, such as bigger clusters more informative or higher precision helps speed up the annotation,

[V] For information on giant component refer here
[VI] For information on course of dimensionality refer here

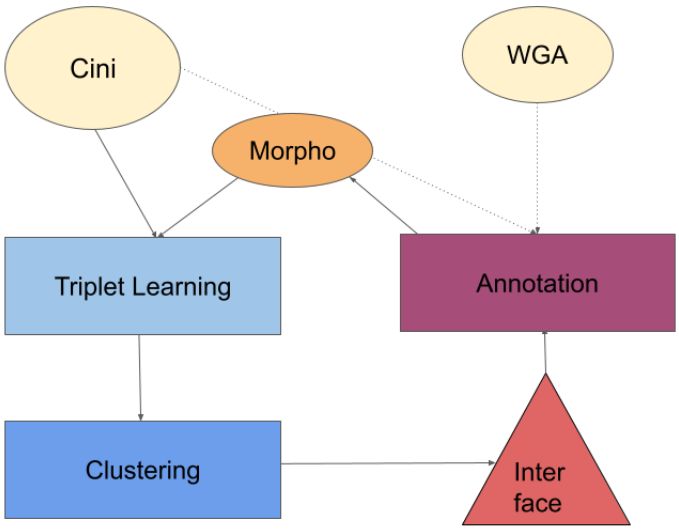


Figure 3.7: Schema of the complete pipeline. On the top, the scheme represents the two data source (Cini and WGA). These were used to annotate the first batch of the morphograph. The morphograph is used to fine tune the model using triplet learning. The descriptors produced with the learning are used for clustering. The clusters are served on a Interface for further annotation. The annotated clusters feed back into the morphograph and begin the cycle again.

are helpful to tune the clustering algorithm.

This process closes the loop of the project (as in Figure 3.7). In fact, we begin with a set of annotated pairs (the morphograph) with the goal of enriching the known patterns with many and valuable new ones. To this end, we train a model on the morphograph, predict on all the images in the collection (obtaining the visual signatures), cluster them using the above methods, annotate them with an online annotation platform (discussed in the chapter Interface) and retrain the original model based on the new annotations. This allows to find desired new annotations at each step and to improve the model to find more. The risk of the iterative learning, however, is double: on one side, one needs to be careful not to cause the model to *diverge*. As a matter of fact, if the re-training occurs only on the newly found connections, the risk of converging to optimal parameters that are only applicable to the new connections is high. This causes an overall divergence of the parameters, which loose the learning of the original training. At the same time, re-training on the full set of data with the new additions is likely to *overfit* the data, obtaining a near perfect precision but never improving its recall.

For these reasons, in our efforts we re-train with a 'half' set. We train on each connection with half the number of sampled c_is ($\frac{N}{2}$) than in the training. For the new connections, we add also the negative connections[VII] in the cluster to the sampled cs. This theoretically yields

[VII] Negative connections are the connections between the newly annotated pairs and the remaining items in the

a learning that is focused on the new areas (especially those where the number of negative connections is high and thus require more learning) but that also does not diverge.

cluster where the annotation comes from. More on this can be found the Interface section

4 Interface

As part of this thesis, a simple web application was developed. The application can be found here. The interface was created with an annotation purpose in mind. In fact, it presented the clusters produced by the model in a way that they could be annotated efficiently. Subsequently, it was expanded to include a view of the morphograph as clusters, where each connected component of the morphograph is represented as a single cluster. Finally, we included a visualisation of the collection as a scalar 2D continuous space. This exhibits the progression inside the collection from one shape to the other, giving an immediate glance into the modification of morphs in art history.

4.0.1 The Replica platform

The interface developed for this thesis can be considered as an extension of the much more complex platform, the Replica platform. The Replica platform was likewise initially developed as an annotation platform with a similar purpose to ours. B. L. A. Seguin, 2018 describes the platform in detail.

A user could search in the collection on the platform with text queries, as an example 'Danae'. The platform would perform a search in the metadata of the collection and return the relevant images. If any pair of these images shared a pattern, these could be annotated and added to the morphograph. If the image the annotator was looking for was not available in the collection, they could upload it to the platform.

After this step of annotation based on the information already recorded in literature, the platform was extended with a basic image search (as in Figure 4.1. The image search was based on the image retrieval process explained above but with the exception that the descriptors were created from the pre-trained model. This step allowed to find and annotate more connections.

The last feature is the map view in Figure 4.2. This consisted of mapping the pre-trained image descriptors to a 2D coordinate space. The images were then visualised in the 2D screen space based on the coordinate of their descriptors. When two images that appeared close to each

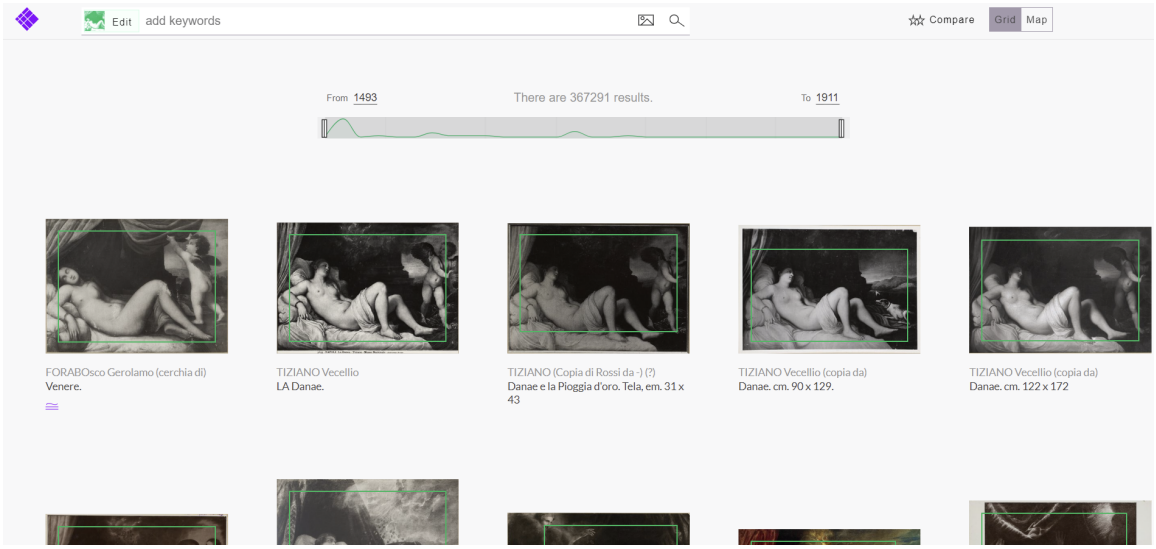


Figure 4.1: Grid view (image retrieval). Retrieved from: Replica. Screenshot of the Replica platform on the diamond.timemachine server showing the grid view results for an image search.

other in the 2D space shared a pattern, these could be annotated.

The annotation tool described in B. L. A. Seguin, 2018 is no longer available online, rather, it evolved into the website linked above. The website presents the same three functionalities of metadata textual search, image search and map view. The last two are now based on the learned descriptors, explained in B. Seguin, 2018, which are a much improved version of the descriptors, being trained for task of pattern propagation detection.

4.1 Clustering annotation tool

Transitioning back to the platform created for this thesis, the first element of the platform is the clustering annotation tool (in Figure 4.3). This was developed to be fit for detailed annotation of groups of images. In fact, clusters annotation differs from a retrieval annotation task. We go in depth into this difference to demonstrate the utility of the clustering approach.

To be able to thoroughly annotate a retrieval model with an annotation tool, one has to query the system with every possible input image and, for each query, annotate any pair of the results shown. If we consider a collection of $C = 100'000$ images and we cut the results of the image retrieval at $N = 20$, the annotation process would require to scan $20x20x100'000$ pairs of images. The process is in $O(N^2C)$. This constitutes a radical improvement from scanning every pair of artworks in the collection as that would amount to $100'000x100'000$ pairs, $O(C^2)$. In the clustering case, however, the annotation time is further improved. We assume that the number of pattern groups in art history is in the order of the thousands. In each cluster, we have a number of images ranging between, more or less, 2 and 20. The annotator has to scan

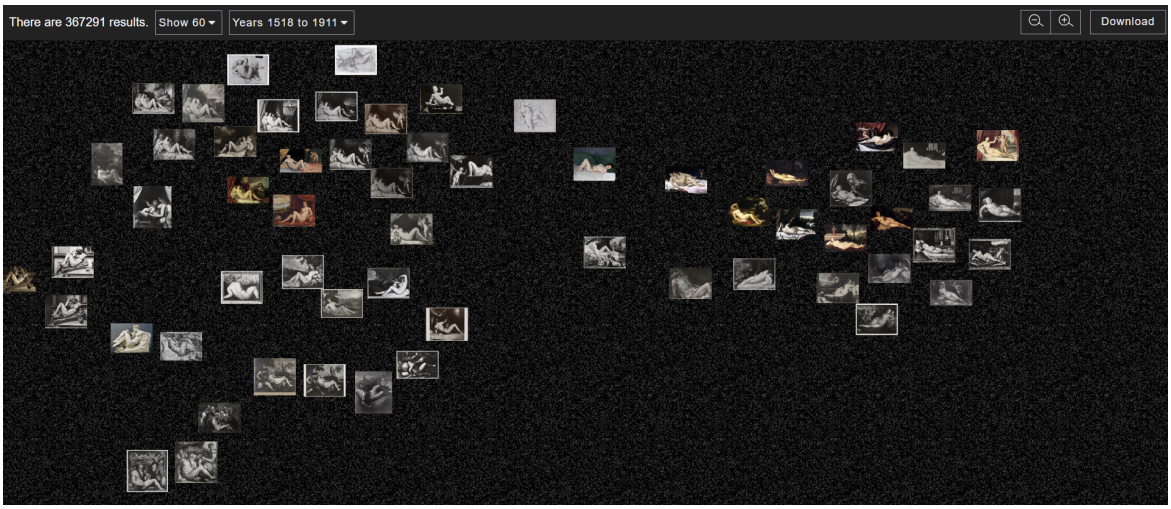


Figure 4.2: Map view. Retrieved from: Replica. Screenshot of the Replica platform on the diamond.timemachine server showing the map view results for an image search.

each group and annotate each pair of images. If we consider the worst case scenario for which we have to scan every pair in each cluster, we have $M = 1'000$ clusters and the cluster size is $S = N = 20$, we have that we annotate $1'000x20x20$ pairs of images ($O(MN^2)$ where $M << C$). While this is a substantial time improvement, we have to note that the potential of the cluster annotation is heavily reliant on the quality of the clusters and, in particular, their recall.

In the discussion above, we assume that the annotator is able to navigate through the clusters in an orderly manner such that they are not presented with the same cluster multiple times or they skip a cluster. Furthermore, the annotator is able to select which set of images in each cluster forms a pattern and which does not. For convenience reasons, the annotator should also be able to indicate that either the whole cluster is correct or the whole cluster is wrong.

The tool was designed with these principles in mind: the annotator is shown one cluster at a time to eliminate latency in loading the images, they can either jump through the clusters at random to obtain a general look on the nature of the clusters (by clicking 'random'), or they can proceed by pressing 'next' to the following cluster. To go directly to a cluster of choice the annotator can input the number of the cluster they wish to explore. Finally, the annotator may search in the metadata (description, author, location, time) to view all the clusters containing at least one match in the metadata.

The annotator has 7 options to annotate each cluster. If the annotator spots a set of images inside the cluster that share the same patter, they can click on that set of images and the button 'confirm pattern'. Upon clicking this button, the server writes into the morphograph, adding a positive edge between every combination of images in the selected set. The annotation writes the time and date of annotation, and also the clustering method and parameters which the annotation comes from. In case the annotator is certain that all the other images in the cluster do not belong to the pattern group of the selected set, they can press the button 'confirm

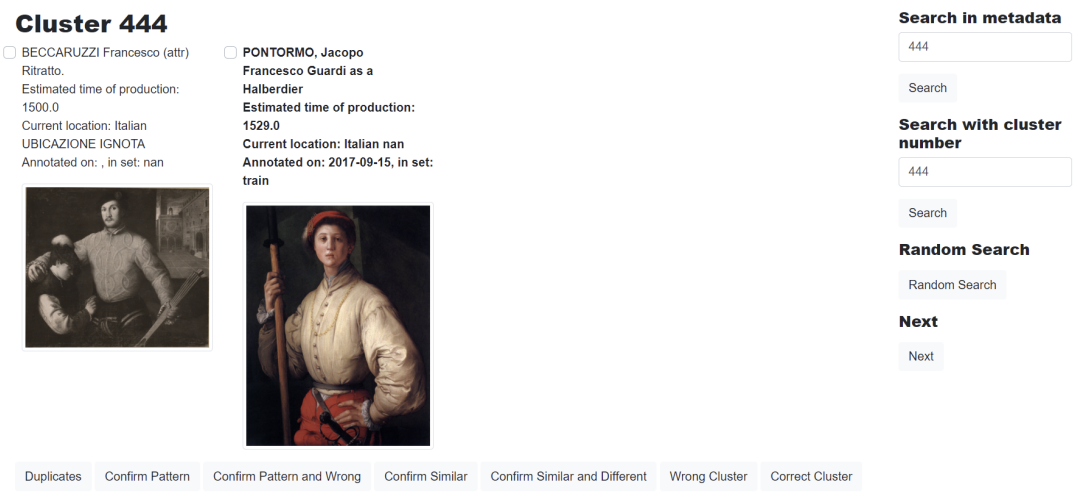


Figure 4.3: Cluster annotation page. Screenshot of the cluster annotation page showing a cluster at random. All the buttons mentioned in the description are visible.

pattern and wrong'. In this case, together with adding the positive edges described above, it adds negative edges between the images in the selected set and all the other images of that cluster that have not been selected. The negative edge indicates that the two images were considered similar by the model (thus placed in the same cluster) but that they are not in reality. The same structure is used to annotate similarity between images, which we consider as a looser connection than pattern (pressing 'confirm similar' or 'confirm similar and different').

Button	Action
Next	Visualise next cluster
Random	Visualise a cluster at random
Search with cluster number	Visualise desired cluster number
Search in metadata	Visualise all clusters with at least one metadata match
Duplicates	Duplicate link between images clicked
Confirm Pattern	Positive link between images clicked
Confirm Pattern and Wrong	Positive link between images clicked and negative with all others in cluster
Confirm Similar	Similar link between images clicked
Confirm Similar and Different	
Similar link between images clicked and different with all others in cluster	
Wrong cluster	Negative link for all images in cluster not in morphograph (and wrong cluster)
Correct cluster	Positive link for all images in cluster not in morphograph (and correct cluster)

Table 4.1: Short cheatsheet of the actions of the buttons on the platform.

Although the collection was deduplicated, we sometimes find that the clusters feature two photos of the same physical element (or identical copies which are indistinguishable). If this is the case, it might mean that the deduplication has failed to detect the duplicate (Seguin 2018), and the annotator can annotate the duplicate link by selecting the two photos and clicking 'duplicate'.

To facilitate and speed up the annotation process, the buttons 'correct cluster' and 'wrong cluster' were added. In the case in which all the images in the cluster belong to the same

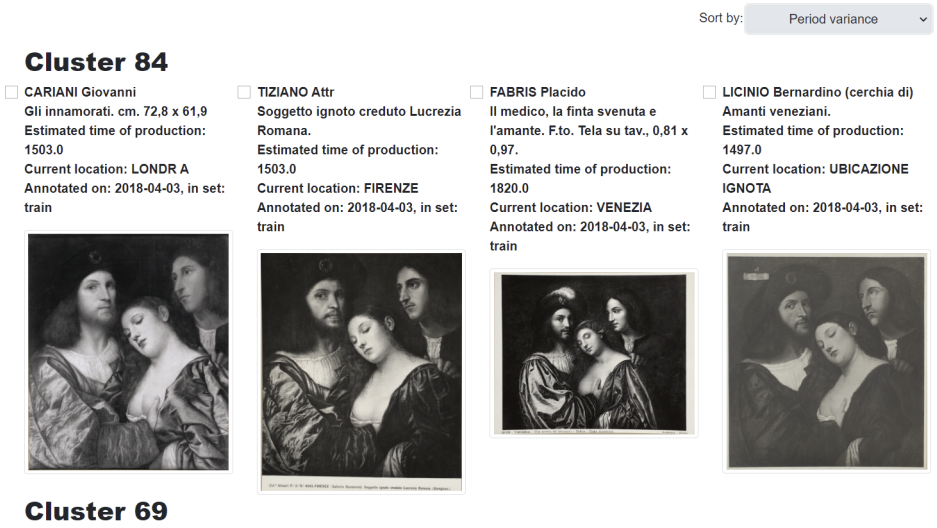


Figure 4.4: Morphograph visualisation. Screenshot of the morphograph visualisation page showing the first cluster when sorting the morphograph groups by 'location variance'.

pattern group (our ideal case), the annotator can simply click 'correct cluster'. This adds a positive edge between all pairs of images in the cluster (unless these were already in the morphograph) and marks the cluster as correct. On the other side, if all of the images in the cluster are wrong, or if all the images that are not yet in the morphograph do not form any pattern with those in the morphograph or among themselves, the annotator can click 'wrong cluster'. In this case, the server adds a negative edge between the images in the morphograph and all others in that cluster and the cluster is marked as wrong.

On the platform, the images that already belong to the morphograph are clearly marked with bold text, while the new images added to the morphograph through annotation are marked with red text. For a short cheatsheet refer to 4.1.

4.2 Morphograph visualisation

A second utility of the platform is the 'morphograph' page (in Figure 4.4). The page presents the morphograph in the most simple way possible: each connected component of the graph[I] is represented as a cluster or group of images. This page is not interactive: all the clusters are shown in the same page, one after the other.

The page gives the annotator the possibility to scroll through the whole morphograph, which, in turn, gives an overview of what was already known in the morphograph and what has been added in the re-annotation process. As for the clusters page, the newly added elements are shown in red.

[I] A connected component of a graph is the maximal set of nodes for which from an arbitrary node in the set there is a path connecting to all other nodes in the set

The page has multiple sorting options. Since the morphograph groups are roughly 400, it is important to be able to sort the groups in the order that is most appropriate to what the annotator is investigating. Therefore, we introduce the possibility to sort the groups by number of elements in the group, number of different authors in each group, different in descriptions of the works (approximated), maximal provenance of the artworks distance and maximal time difference between when the artworks were made.

In particular, the sorting option named *description variance* is a very loose approximation of results with respect to iconography. We use the descriptions, that have all been automatically translated to English, removed of noisy information (i.e. size of painting), and clustered with `kmeans++` text clustering on Manhattan distance. The variance in the descriptors is then evaluated as the number of different textual clusters. The *author variance*, on the other side, was computed from the Author field that has been refined by B. Seguin, 2018, using only the last name provided. We consider members of the same family to not introduce any author variance. To find patterns propagating in workshops and through copyists, we introduce a sorting by *artist attribution*, which counts the number of different attributions types in the same cluster. The count is obtained using the non-processed artist name, which contains specifications such as '(scuola), (bottega di)'. These were extracted and clustered as above to remove unnecessary overhead, and the number of different textual clusters was counted to obtain the variance. Finally, the *variance in time and space* were computed as the maximal variances in the clusters. In the case of the time, the clusters were ordered by maximal year of production difference, while for place, they were ordered by maximal longitude and latitude of artist death place distance.

Although these metrics can be indicative, the imprecise nature of the metadata makes the results only very rough approximations and are to be considered as merely a first attempt to open a future avenue of research.

4.3 Visual clustering

While the above pages are effective for annotation and single cluster views, they suffer by two main limitations: they do not offer a *global view* of the images and they *discretise to unit distance* the space. In fact, the shape of art history can be considered as a gradual evolution from one pattern to another whilst, in the pages described above, a pattern group can only be seen in isolation, without its relation to other pattern groups nor in its gradual transformation within the cluster. For this reason, we included a third page visualising the pattern migration space in a 2D interactive space.

Attempting to recreate a morphological space of the history of art pre-dates the birth of computers. Aby Warburg in the 1920s, assembled the already mentioned Bilderatlas Mnemosyne over the course of several years and modifications (Diers et al., 1995). In the atlas, the art historian breaks the linearity of the traditional books, transposing the space of art into 2D black boards, the panels. He traces recurring visual connections of Pathosformel across the

Figure 4.5: Mnemosune Atlas. Retrieved from: artishock.The image is from the installation view at Haus der Kulturen der Welt, Berlin of the Bilderatlas.

millennia of the history of art. He included statues, paintings, cards, photographs from antiquity to the Renaissance and to contemporary works and collated them into panels with a spatial meaning (Didi-Huberman, 1996). *(more on Warburg)*

Warburg's methodology set new standards: he opened the art world to rearranging canonised images and looking at them across epochs while traversing the boundaries between art history and numerous other modern disciplines, such as psychology, and cultural studies.

A century later, the value of giving spatial coordinates to images is being explored by many digital projects. Google Arts developed the t-SNE Map experiment, DHlab Yale developed PixPlot and Replica project included the map view recently explained (Diagne and Barradeau, n.d.; Duhaime, 2019; B. Seguin, 2019.

With t-SNE Map, one can explore an interactive 3D landscape automatically created from images of artworks organised by they similarity. t-SNE was used to reduce the dimensionality of the feature descriptor of the images to three, so that it could be visualised in the pseudo-3D screen space. Most of the visualisation uses three.js and webGL. The landscape, visible in Figure 4.6, is attractive, but the similarities of the images are not fine-grained enough for our purposes and the navigation in the 3D space from a 2D screen is not natural (Diagne and Barradeau, n.d.).

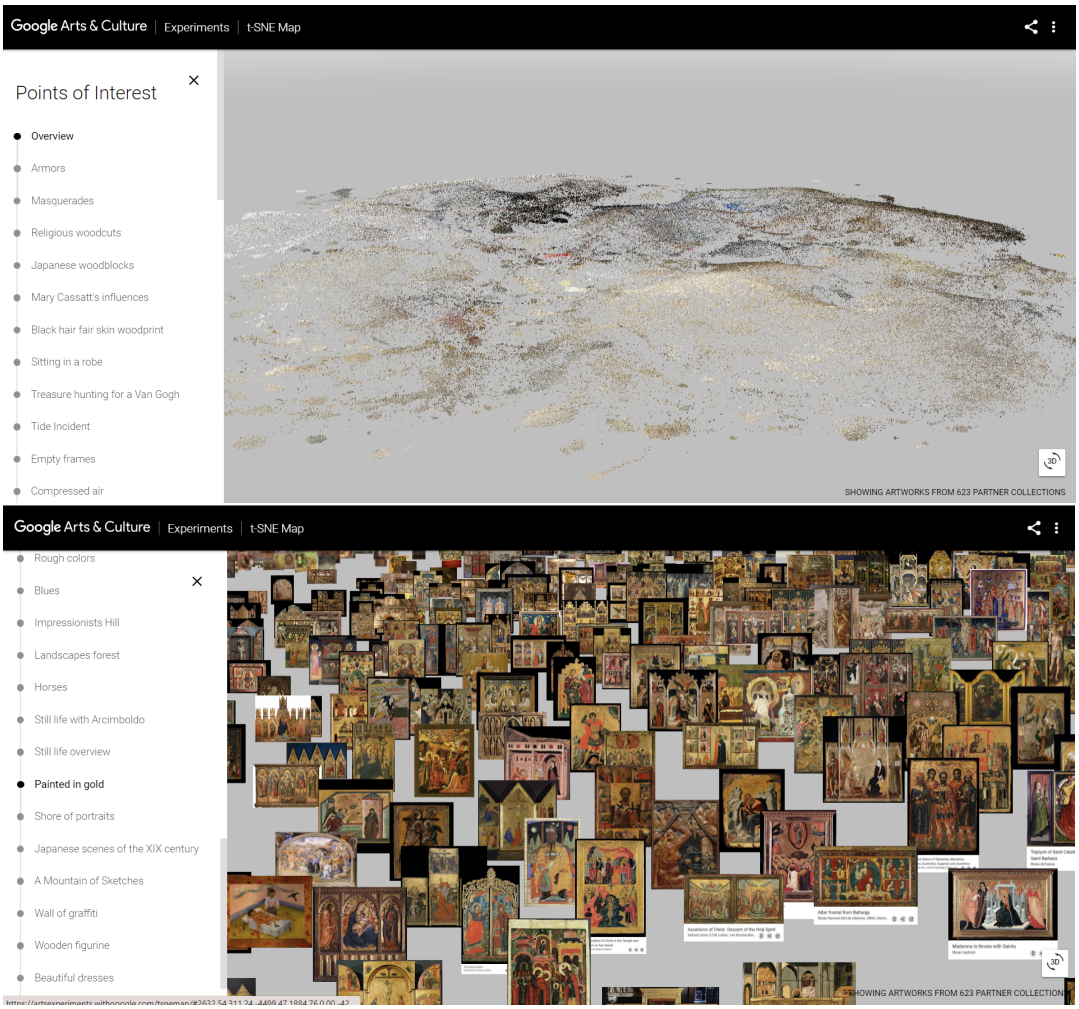


Figure 4.6: Google Art t-SNE map (close-up on the bottom). Retrieved from: t-SNE map. The images are screenshots of Google Art and Culture's t-SNE map, showing images in a natural 3D landscape. The site allows guided navigation that redirect to desired clustered areas, as in the case on the bottom, to golden paintings.

PixPlot represents a reworking of the previous Google arts project, where they include over 27,000 images from the nineteenth-century Meserve-Kunhardt Collection and project them into a two-dimensional manifold with the UMAP, rather than t-SNE. The visualisation is then made pseudo-3D to avoid occlusions. In this case, being the projection to a 2D space, the navigation is eased. Once again, the general clusters appear well defined but contain little fine grained information (Duhaime, 2019).

With B. Seguin, 2018's map view, the projection is to 2 dimensions which are forced to be non-overlapping to remove occlusions. The fine-grained information is well represented, but the 2D non-overlapping constraint limits the number of possible images shown simultaneously to only few hundreds.

Figure 4.7: DHYale PixPlot (close-up on the bottom). Retrieved from: PixPlot. The screenshot shows similar functionalities to t-SNE. One immediate difference is the presence of delineated clusters which were created with UMAP and could not be created by t-SNE.

The visualisation created for this thesis integrates the `pixplot` framework of scalar pseudo 3D visualisations to the fine-grained information contained in the descriptors learned after B. Seguin, 2018. We adapted the library released by DHlab Yale for work for our case. The final result can be seen in Figure 4.8. The result of the visualisation is promising: the clusters can be easily detected, the 2D location is meaningful, and the page is able to load the images of the subset without any latency. In addition, the `pixplot` library offers other views: using UMAP to evidence the location of the clusters, sorting alphabetically, in time, filtering. The library offers the option to filter by category, thus, we used this option to incorporate a filtering with what is already in the morphograph and where it is.

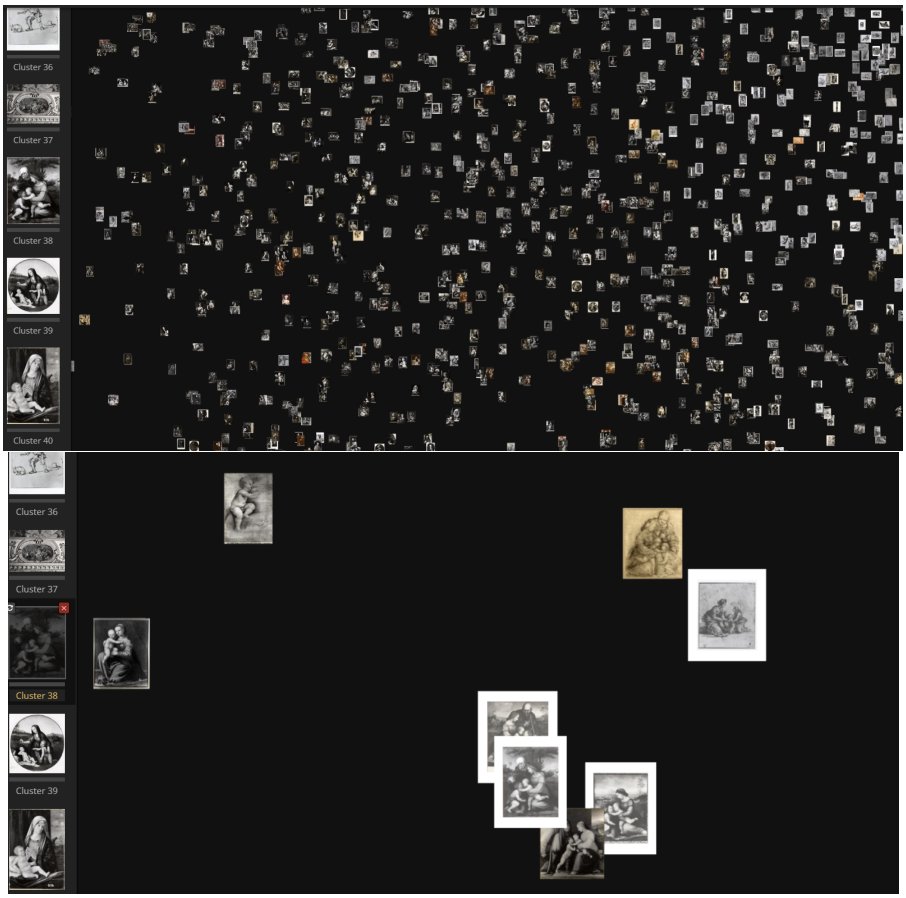


Figure 4.8: Visual clustering (close-up on the bottom). The screenshots show the result of Pixplot adapted for this task. It utilises the specialised visual descriptors learned with triplet learning, the clustering with OPTICS (clusters are visible on the left and can be used for navigation) and the 2D coordinates computed for this thesis with t-SNE.

5 Results

This section presents the results of the feature descriptor learning and the clustering efforts, introducing the experimental setup and the metrics that were selected and created for the tasks. In particular, the feature descriptor learning is evaluated in terms of its performance at the image retrieval task, while the clusters are evaluated both with respect to their coherence to the morphograph and the annotations based thereof. The iterative update is accounted for in both sections.

5.1 Feature learning

5.1.1 Experimental setup

In our experiments, we use the `Pytorch 1.11` deep learning framework. We pre-process the images to the fixed size of $320x320x3$ as input shape, normalise it across the three channels to mean of $0.485, 0.456, 0.406$ and standard deviation of $0.229, 0.224, 0.225$ (which is the standard in computer vision) and pass it to the model with a batch size of 4. The batch size and resolution could not be increased due to memory constraints of the GPU.

The input images were augmented as follows: the images were originally resized to $420x420x3$ size, cropped randomly to the final size of $320x320x3$, passed through a colour jitterer (in brightness, saturation and contrast with a probability of $p = 0.3$), a random horizontal flip with the same probability and a small 5 random rotation. We experiment both with and without this augmentation. The augmentation steps were selected as they mirror most of the invariances that we wish our model to learn.

The ground-truth data is split into training, validation and test set using the connected components of the morphograph. In fact, the python `networkx` library was used to create an undirected graph representation from the set of edges of the morphograph and to extract its connected components. All the artworks belonging to the same connected component were assigned to the same set. This prevents the model from learning and predicting on the same

pattern group. Two thirds of the connected components were included in the training set, one sixth to the validation and the remaining sixth to the test. The resulting training set contains 4513 pairs of artworks, the validation 1460 and the test 1311.

The Adam optimiser was used with learning rate $10e^6$, and a learning rate scheduler of $10e^2$ decay every 1000 steps. An L2 weight decay of $10e^5$ was included. We retrain all the layers of the model except for the batch normalisation layers which remain frozen. The Hinge loss margin was set to $m = 0.01$ and the maximal intra-cluster distance to $m_2 = 0.13$. We experimented both with an without the additions of the intra-cluster constraint and anchor swap to the loss. We ran all the models for 30 epochs and store the model with lowest validation loss or best image retrieval performance (using early stopping).

To create the training set, we select $N = 5$ c_is for each pair of a_i and b_i. We therefore train on $4513x5$ training triplets and validate on $1460x5$ triplets. The training and validation set are updated after every epoch according to the iterative hard negatives sampling procedure described in the Methodology.

Retraining set-up

After training the model for 30 epochs with early stopping, we predict on every image in the collection and store the feature descriptors (8959 vectors of shape $1x2048$). We cluster the descriptors with one or more methods and annotate all the clusters with the annotation tool described in the Interface section. We obtain a list of *positive* and *negative* edges for each cluster after the re-annotation.

We update the morphograph by adding the newly annotated *positive* pairs. When one of the two images in the pair belongs to an already existing pattern group, we add the new image to the same set as that group. When the pair is completely new, we add this to the training set.

Based on the annotation, we create a new training set. For each image a_i in the updated morphograph, we sample a b_i from the $N(a_i)$ of the updated morphograph. To obtain the c_is we follow two cases. If we added a negative edge between that image and others during the annotation, we consider these other images as c_is and sample one other image with the hard negative sampling discussed in the Methodology. When no negative edge is added, we sample $N = 3$ c_is with hard negative sampling. In the first case we have a variable number of training triplets for each (a_i, b_i) pair (usually above 3), in the second case we have 3 for each pair.

We retrain with the new set for an extra epoch from the last stored model and repeat the process. This method allows us to consolidate and learn from the new patterns and to simultaneously inform the model of the mistakes in the cluster assignment.

5.1.2 Evaluation

To evaluate the performance of the model we assess its results at the image retrieval task. We use the images of the validation and test set as queries to evaluate the results.

The ordered lists that are outputted for every image are evaluated using the standard image retrieval metrics. These include: mean average precision (MaP), and recall (R) at different levels, in this case 20, 50, 100, 200 and 400. The first metric captures how soon the relevant documents are retrieved, the latter quantifies the number of relevant images that are retrieved if we cut the results at 20, ..., 400.

To compute these metrics, the ordered list of retrieved images is converted into a binary list of zeros and ones depending on whether the image at position j is a neighbour of the input image (in $N(a_i)$) in the morphograph or not (i.e. if there is an edge between the input and the retrieval image at j). The MaP is computed from the full signal (which, for computational reasons has been cut to 4000), while the recall is calculated as the number of ones in signal cut at different thresholds over the total number of neighbouring nodes.

In addition to these metrics, we include some indicators on the positions of the ground-truth images in the binary list. We cut the resulting signal to the first 400 and compute the mean minimum position (the average first position in which a one is found), the mean median position (the average of the median of all the positions of the ones for each query) and the mean position of all the neighbours of the image in the morphograph.

These metrics have been chosen to give a comprehensive understanding of the performance of the model and are implemented on the basis of the code here.

We evaluate the performance of the re-training with the same metrics. In addition, we keep track of the growth of the morphograph after each annotation session. We compute the number of new positive pairs, new images, what portion of these images have already are linked to an image already in the morphograph (and would therefore extend already existing pattern groups), the number of pattern groups that receive at least one new image, the portion of pairs that consist of all two new images (in number of new images) and the number of new groups these create.

5.1.3 Model comparison

In the Methodology, we discussed a number of modifications to the original Replica model. We test in this section how the modifications perform and subsequently discuss the training of the final model.

We experiment with a number of variations of the ResNet, ResNeXt and EfficientNet models as reported in Table 5.1 and evaluate the performance of pre-trained architectures. We can observe that, while ResNet still achieves a considerably good performance, the more recent

Architecture	pooling	resolution	mean position	mean min position	mean median position	map	recall at 400	recall at 200	recall at 100	recall at 50	recall at 20
resnet50	avg	240	335.41	174.97	376.13	0.06	0.39	0.29	0.22	0.18	0.12
resnet50	avg	480	328.47	168.09	369.30	0.07	0.43	0.33	0.26	0.19	0.13
resnet50	max	240	354.26	218.47	388.97	0.03	0.28	0.21	0.15	0.11	0.07
resnet50	max	480	355.20	217.79	389.54	0.03	0.30	0.21	0.15	0.11	0.08
resnet101	avg	480	305.65	134.42	353.64	0.07	0.46	0.35	0.27	0.21	0.16
resnet152	avg	480	302.47	133.67	349.42	0.09	0.48	0.37	0.28	0.22	0.17
efficientnet7	avg	240	382.60	305.78	398.53	0.01	0.15	0.08	0.05	0.03	0.01
efficientnet7	avg	480	375.39	287.90	396.04	0.01	0.18	0.10	0.06	0.03	0.02
efficientnet7	max	240	374.26	272.03	397.76	0.01	0.18	0.11	0.07	0.04	0.02
efficientnet7	max	480	370.97	273.71	392.86	0.01	0.19	0.12	0.08	0.05	0.03
resnext-101	avg	480	**285.85**	**119.34**	**327.06**	**0.12**	**0.54**	**0.43**	**0.35**	**0.26**	**0.20**

Table 5.1: Pre-trained architecture, resolution and pooling method comparison.

model of ResNeXt performs best, perhaps due to the impact of the cardinality dimension in the feature extraction. The poor performance of EfficientNet is surprising. We believe that it can be explained by the insufficient hyper-parameter search for the best terminal layer of the model. In fact, we removed the last layer of the model, while we might obtain improved results by selecting different layers. Overall, we see that increasing the resolution always leads to an improvement in performance and that average pooling seems to outperform the maximum.

We therefore proceed with ResNeXt with average global pooling in the following experiments. We could not fit a resolution of 480 in the GPU memory so we decreased it to 320. After having chosen the resolution, architecture and pooling, we test the performance of the modified loss in comparison to the simple triplet loss and to the pre-trained model baseline. We observe in Table 5.2 a significant improvement with respect to other two, with the most significant improvement in the MAP.

Modification	mean position	mean min position	mean median position	map	recall at 400	recall at 200	recall at 100	recall at 50	recall at 20
baseline	287.44	126.27	338.56	0.13	0.53	0.43	0.33	0.26	0.19
simple loss	223.40	65.36	232.07	0.34	0.75	0.69	0.62	0.55	0.45
intra loss + anchor swap	**199.73**	65.40	**204.39**	**0.41**	0.78	0.72	**0.66**	**0.59**	**0.50**
intra loss + anchor swap + retrain 1	202.41	62.81	205.28	0.40	0.77	**0.73**	0.65	0.59	0.49
intra loss + anchor swap + retrain 2	205.13	**61.49**	209.37	0.41	**0.78**	0.72	0.65	0.58	0.49

Table 5.2: Table with results of the best efforts

In the same table we also find the effect of the two retraining steps. In the first step of annotation, we include 222 new pairs, with a total of 167 new images, 91 of which enrich 26 existing patterns, and 76 which create 36 entirely new pattern groups. This causes an overall increase of 2.84% in the size of the morphograph. In the Table 5.2 we see that the retraining causes a very small variation, with an improved score only in the recall at 200.

In the second step, we include 126 new pairs, with a total of 91 new images, 52 of which enrich 17 existing patterns, and 39 which create 17 entirely new pattern groups. This causes an overall increase of 1.61% in the size of the morphograph. The second retrain causes a slight improvement in min position and recall at 400.

While the improvement in the score is negligible, we see that continuing training on the model after the 26^{th} epoch causes a noticeable decrease in performance, and this is alleviated by the retraining. From Figure 5.3 and Figure 5.4, we immediately notice the considerable

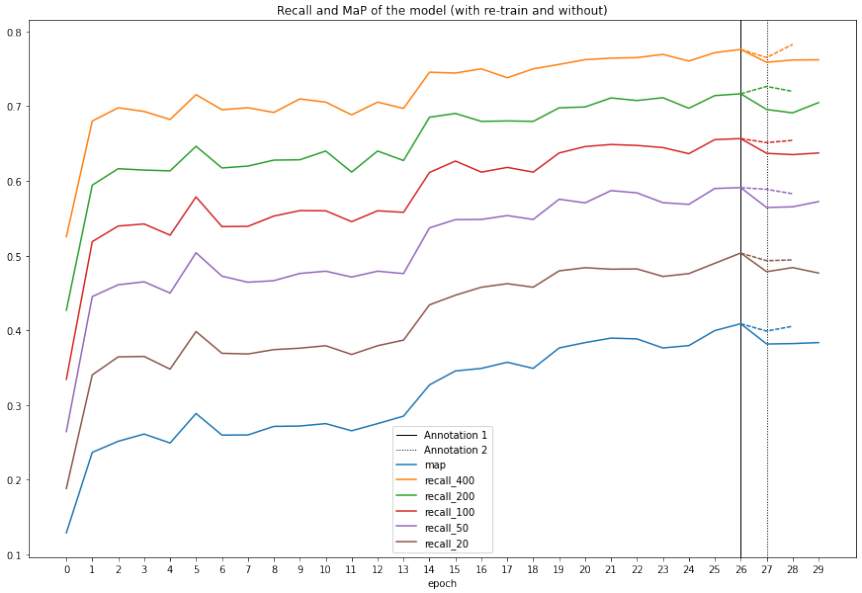


Figure 5.1: Recall and MaP scores per epoch of the final model. The vertical lines indicate the annotations, the learning bifurcates in retraining (dotted) and continued training (continuous line). We observe how the continued training start to drop the performance, while the re-training alleviates the trend.

improvement over the pre-trained model. We see, additionally, that the minimum positions on the train set have almost converged to 0, while this is not yet the case for the validation and test set. This is an indication that the model is overfitting the data.

5.1.4 Error analysis and discussion

We take into consideration a few interesting examples to assess qualitatively the learning of the model. We expect ResNeXt pre-trained for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) to perform considerably well on realistic paintings, frescoes, statues (and more). In fact, the model has been trained on the huge corpus of images of ImageNet, containing naturalist and urban environments, animals and objects. The model is aware of perspective changes to realistic images, and, due to the cardinality of ResNeXt, possibly also to difference in scale.

We see indeed, in the first example in Figure 5.5, that the pre-trained model performs rather well on the detection of Palazzo Ducale and Piazza San Marco series. This series, as discussed earlier, is rather naturalistic, inducing only slight perspective changes. Interestingly, after training in Figure 5.6, we observe that even depictions from completely different perspectives (The Bucintoro and The Doge Palace) of the same building were retrieved among the 11 most similar images. This indicates a rather profound invariance to perspective changes and a remarkable capacity to detect the object constituting the pattern, in this case, the Palazzo

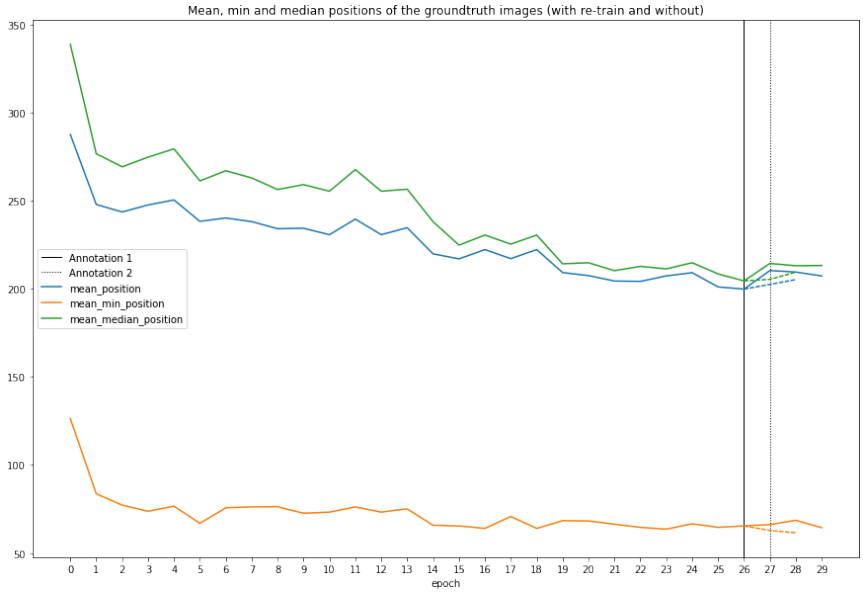


Figure 5.2: Min, mean and median positions per epoch of the final model. The same trend can be observed with these metrics.

Ducale. In effect, the model is not distracted by differences in the sky nor by the different boats and activities linked to the setting.

The second example, whose pre-trained results are visible in Figure 5.7, illustrates a lack of invariance to medium of the pre-trained model. When querying the model with Gambara's Deposizione nel Sepolcro, the results do not share any compositional nor contenutistic resemblance (with the exception of Caracca's Cristo morto, la Madonna e due Angeli). These are rather dominated by the same colours (grays) and are all drawn with pen or pencil. After fine-tuning, the model acquires great stylistic invariance, retrieving numerous pattern images that were not in the morphograph at the time of training. In particular, it remarkably finds even Bassano's chalk depiction. Furthermore, it is capable of accounting for the addition and removal of secondary figures and changes in the background. A similar learning dynamic can be individuated in Michelangelo's Flagellation of Christ, a rather rough sanguine drawing (in Figures 5.9, 5.10) that could not be retrieved using the pre-trained descriptors but could be successfully retrieved with the fine-tuned ones.

While we have thus far considered examples of queries in the training set which exhibit a virtually perfect fine-tuned performance, we now look into two cases which were not included in the training. The first is the notorious Tiziano's Venus with a Mirror. In the pre-trained results in Figure 5.11, we reckon that the model is paying attention to artworks with half or full body female figures, with rather warm colours. It is, however, not aware of the importance of the position of the woman. In Figure 5.12, the model exhibits a much increased awareness to the pose, that translates into findings such as Vasari's Allegoria della Castità and Tiziano's Woman in a Fur coat, where the resemblance is circumscribed to the position of the arm.

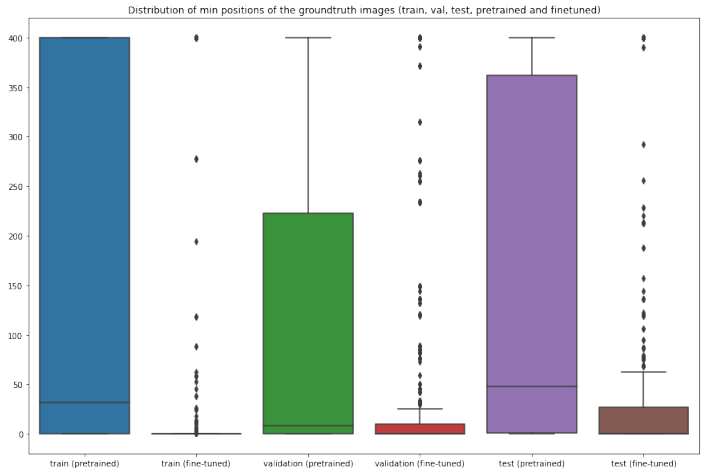


Figure 5.3: Minimum positions of the pretrained and finetuned best model (retrain 2) for the train, validation and test set. The figure shows the effect of learning (each pair have the first block representing the pre-trained and the second the fine-tuned). Furthermore, each block represents a set, we observe that the first block, the training set, obtains an almost zero min position distribution, indicating overfitting.

We observe that, differently from the training examples, even the fine-tuned results are not impeccable.

The last case, in Figures 5.13, 5.14, demonstrates an extremely complex example for this level of feature learning: invariance to the contentHere by content we mean semantic understanding of the scene, which is a typical feature of CNN-based models. In fact, while the pre-trained case does not show any pattern, we observe that the fine-tuned descriptors yield a number of other images featuring breastfeeding, a similar action to that of the query image. The model, however, fails to find the images that share a pattern with the query, such as the example in Figure 5.15. This is probably due to the completely different figurative content between the images. Another, more problematic but less likely reasons, it that the model poses the attention on the figure on the left rather than the woman. This may be indicative of a small bias that was created by the morphograph, which trains the model with disproportionately more examples in which the pattern is to be found on the pose of the Child than that of the Madonna.

Beyond these examples, we observe a clear trend of overfitting. During the experiments, a great attention was dedicated to regularising the model and preventing the phenomenon (with increased margin, weight decay, augmentation, lower learning rate, fewer epochs). This phenomenon, however, seems unavoidable. We believe that overfitting is induced, in our experiments, by three factors. The first is *inherent to the training procedure*. If we take in example Michelangelo's Flagellation of Christ (Figure 5.9), the image has 7 neighbours in the morphograph. Knowing that every image a_i is trained with every possible b_i in the neighbourhood of $N(a_i)$ and for each of the combinations (a_i, b_i) with $N = 5$ c_is, we reckon

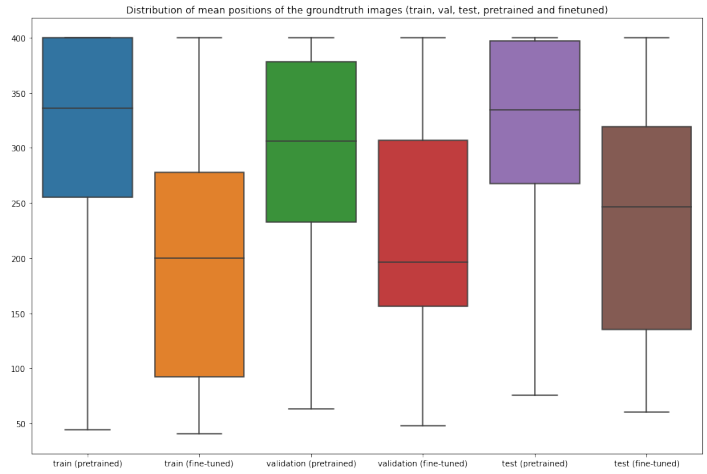


Figure 5.4: Mean positions of the pre-trained and fine-tuned best model (retrain 2) for the train, validation and test set. Similar results can be observed as for the figure above. We do not observe the same intensity of ovefitting.

that the Flagellation of Christ appears in the training $7x5 = 35$ times every epoch rather than the usual one time in standard classification models. Secondly, it is trained with *extremely similar bs and cs*. This inevitably encourages the model to fit the image very thoroughly, thus leading to overfitting. Finally, the task itself is an exceptional case for CNNs. While these are usually adopted for semantic level understanding of scenes, we are forcing the CNN to *learn purely visual and detailed signals* of the artwork, which, again, requires very well tuned parameters.

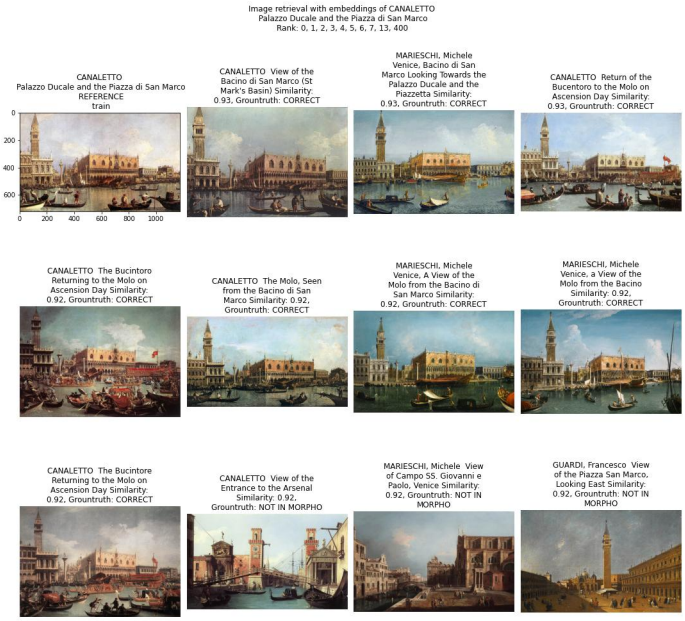


Figure 5.5: Canaletto, Palazzo Ducale and the Piazza di San Marco. Retrieval with pre-trained descriptors. We observe a rather correct retrieval, except for the last three, representing different places.

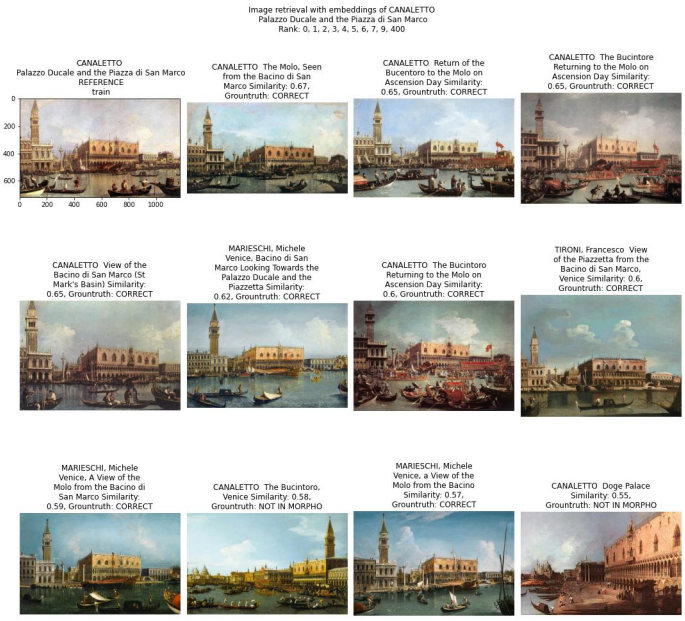


Figure 5.6: Canaletto, Palazzo Ducale and the Piazza di San Marco. Retrieval with fine-tuned descriptors. We observe a perfect retrieval.

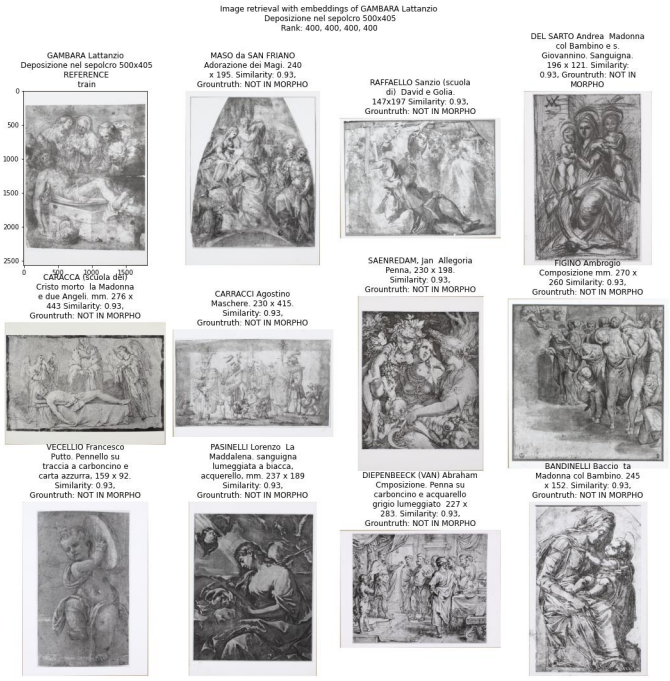


Figure 5.7: Gambara, Lattanzio Deposizione nel sepolcro. Retrieval with pre-trained descriptors.

Figure 5.8: Gambara, Lattanzio Deposizione nel sepolcro. Retreival with fine-tuned descriptors

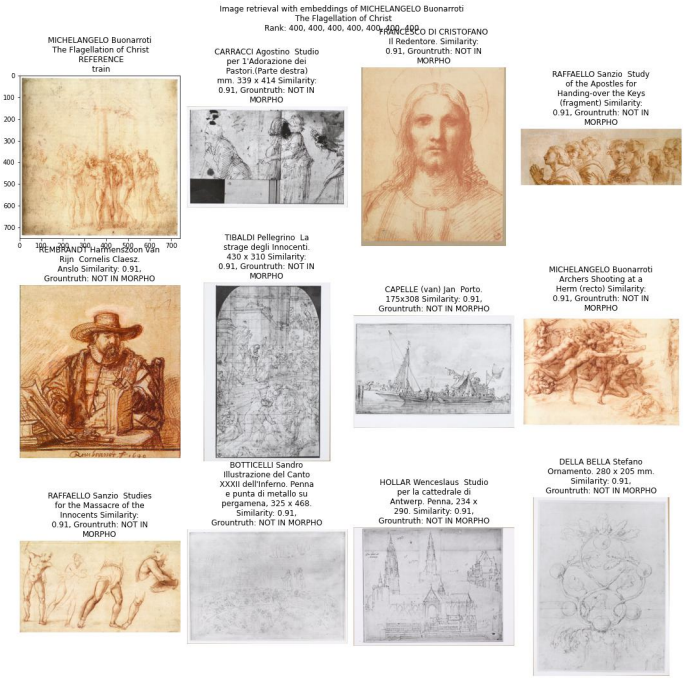


Figure 5.9: Michelangelo Buonarroti, The Flagellation of Christ. Retrieval with pre-trained descriptors.

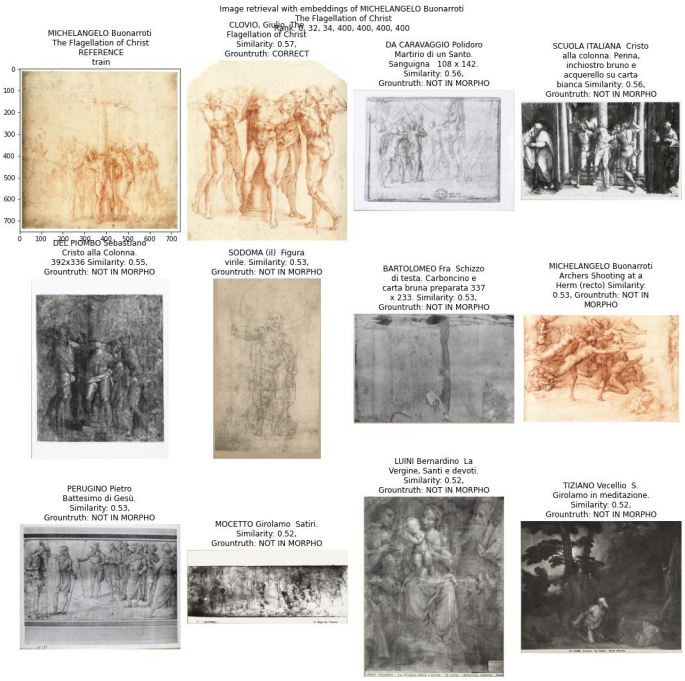


Figure 5.10: Michelangelo Buonarroti, The Flagellation of Christ. Retrieval with fine-tuned descriptors.

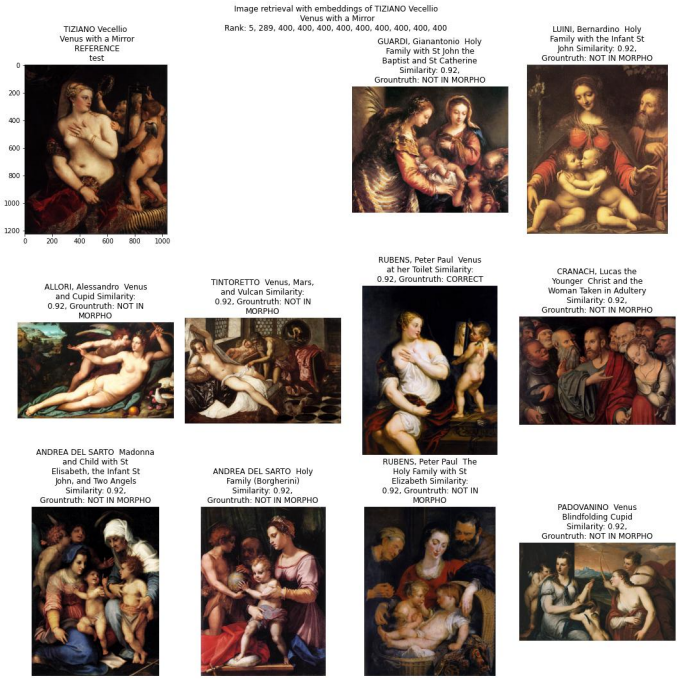


Figure 5.11: Tiziano Vecellio, Venus with a Mirror. Retrieval with pre-trained descriptors.

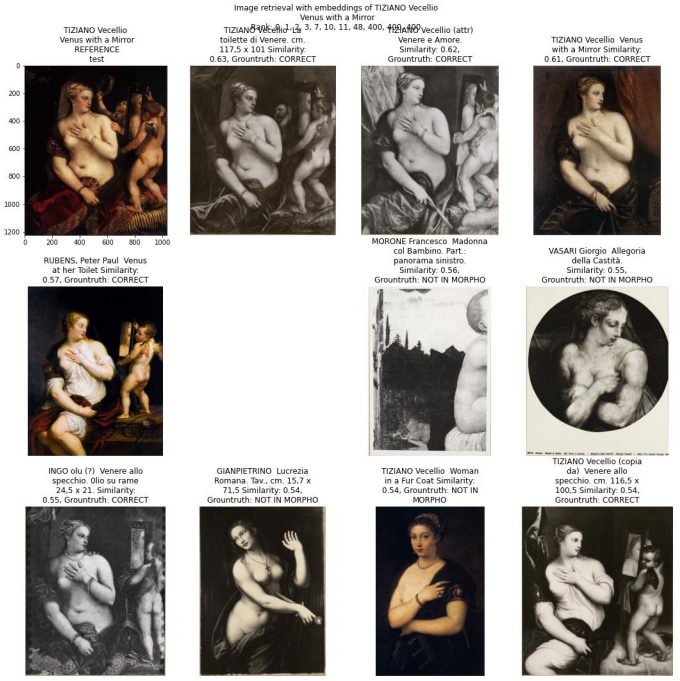


Figure 5.12: Tiziano Vecellio, Venus with a Mirror. Retrieval with fine-tuned descriptors.

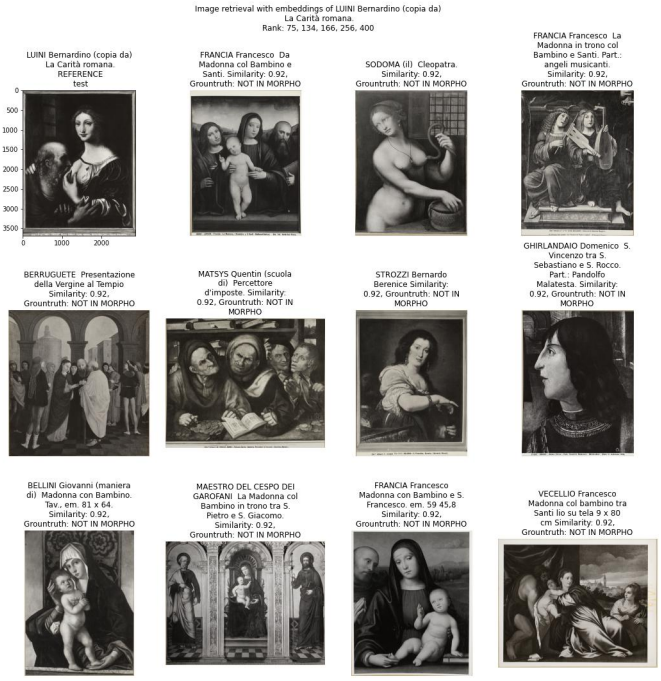


Figure 5.13: Luini Bernardino (copia da), La Carità romana. Retrieval with pre-trained descriptors.

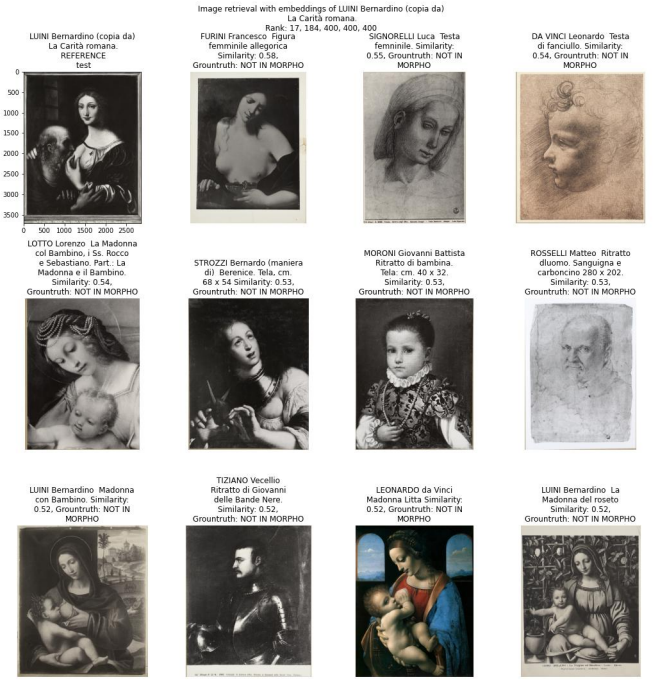


Figure 5.14: Luini Bernardino (copia da), La Carità romana. Retrieval with fine-tuned descriptors.

Figure 5.15: Luini Bernardino, Madonna. Neighbour of Luini Bernardino (copia da), La Carità romana.

5.2 Feature clustering

5.2.1 Experimental setup

To perform image clustering, we rely on the python machine learning library `sklearn`. The predictions of the model on the full collection (8959 vectors of shape $1x2048$) are used as input to the clustering. We recall that, of these 8959 images, around 1800 already belong to the morphograph, while the remaining do not have any annotation attached.

We perform clustering as a series of three steps, the first two being optional. First, we *reduce the dimensionality* of the input to 500 dimensions using `sklearn`'s Principal Component Analysis (PCA) or to two dimensions using `sklearn`'s t-SNE. We experiment with dimensionality reduction to avoid the curse of dimensionality that debilitates many clustering algorithms. Subsequently, we *remove the outliers* using `sklearn`'s DBSCAN outliers detection[I] to account for the fact that most images will not belong to any pattern group. Finally, we *cluster* the resulting image descriptors using k-means, DBSCAN or OPTICS.

We experiment with different configurations of the three steps and different parameters for the clustering. In particular, we focus on tuning the *number of clusters* for K-means and the *maximum distance* between two samples for DBSCAN and OPTICS. We note that, given the maximal intra-sample distance in the loss, this value is upper bounded (and approximated) by $m_2 = 0.13$. We force DBSCAN and OPTICS to include at least 2 samples per cluster and count as outliers every singleton cluster. When the choice was possible (in DBSCAN and OPTICS), we used cosine distance[II] for the clustering. We ran k-means for 100 maximum iterations and with 10 random initialisations.

5.2.2 Evaluation

Being an unsurpervised task, image clustering does not have standard supervised evaluation metrics. The metrics, in fact, usually rely on the coherence within the clusters and the incoherence with other clusters. In our case, we can take advantage of the information in the morphograph and the annotation for a semi-supervised evaluation.

We envisage the morphograph as a set of clusters, as for the second page of the Interface. We represent each connected component of the morphograph as a self-standing cluster. To compute the precision and recall of the clusters by the model, we compare the morphograph clusters to the ones at hand. For clarity we refer to the clusters in the morphograph as *groups* while the clusters of the model being evaluated are referred to as *clusters*.

For a group in the morphograph, if at least one image of the group is not an outlier, we define as *cluster precision* the maximum number of images in a cluster that belong to that group in

[I] The clustering method places images in cluster -1 when these are considered outliers.
[II] Cosine distance is defined as 1 - cosine similarity

the morphograph over the number of images in the cluster. To account for the fact that we constrain to only cases where it retrieves at least one element, we subtract one from the top and bottom side of the division. The cluster precision is:

$$\hat{k} = \arg\max_{k \in C}[size(g \cap k)]$$

$$P_g = \frac{size(g \cap \hat{k}) - 1}{size(\hat{k}) - 1} \mid size(g \cap \hat{k}) > 0$$

where g is the group for which we are computing the precision, C are the clusters and k is the cluster with maximal intercept with g.

The *mean cluster precision* is the average cluster precision over all groups in the morphograph, divided only by the number of groups $size(\mathcal{M}) - X$ that had at least one intersection with the clusters, as:

$$X = \sum_{g \in \mathcal{M}} 1 \mid size(g \cap \hat{k}) = 0$$

$$P = \frac{\sum_{g \in \mathcal{M}} P_g}{size(\mathcal{M}) - X}$$

The mean cluster precision encapsulates how precise the clustering is in the cases where the images of the morphograph are not considered outliers. This metric can be referred to also as the purity of the clusters.

Differently, we define the *cluster recall* for every group in morphograph to be 0 if all the images in the group were classified as outliers. If at least one image is not an outlier, we define the recall as the ratio between the maximum number of images of the group that are clustered together over the total number of images in the group, as:

$$R_g = \frac{size(g \cap \hat{k})}{size(g)}$$

$$R = \frac{\sum_{g \in \mathcal{M}} R_g}{size(\mathcal{M})}$$

This recall metric is rather strict as it penalises the clustering even in cases where the images are clustered but not perfect.

Additionally we keep track of the number of images that are clustered (those that are not marked as outliers) and the number of clusters obtained. Ideally, we would like to cluster as many images as possible in as many meaningful clusters as possible.

Finally, we take advantage of the annotation of 'CORRECT' and 'WRONG' cluster to compute the ratio-accuracy of the clustering experiment as the ratio of correct clusters over the wrong clusters. Being the annotation an extremely time consuming task, this metric cannot be

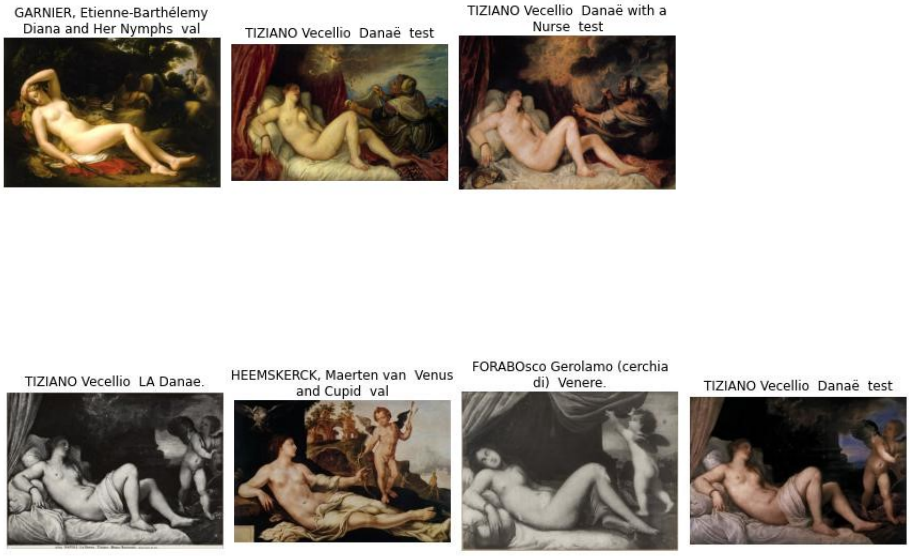


Figure 5.16: Danaë with a Nurse cluster. Obtained with k-means.

computed on all clustering efforts, rather only the fully annotated ones.

5.2.3 Clusters comparison

We present the results of the clustering efforts in Table 5.3. One can immediately see that the results are not very clear cut, nor favouring a specific method. The first observation that can be derived is that, while the precision appears very high in almost all models (except the one not performing outlier detection), the recall is low for all models. K-means with outlier detection generally obtains the highest recall and worst precision. DBSCAN, specularly, has very high precision but low recall.

We notice, in addition, that a rather small portion of the original images is retained after outlier removal, and the amount of clusters is generally about 2 times the number of clusters in the original morphograph. Interestingly, in the clustering with the pre-trained descriptors, we observe that very few data points can be reasonably connected into clusters, while this increases with training and re-training, demonstrating the usefulness of the learning.

Furthermore, while the results for DBSCAN are very heavily influenced by the maximum distance (ϵ parameter), OPTICS is a more robust method. Along the same lines, DBSCAN easily agglomerates most of the images into one giant component[III] in cluster 0, while this

[III] The term giant component comes from network analysis, indicating a cluster that is many orders of magnitude larger than the others that typically forms in real-life networks

Figure 5.17: Danaë with a Nurse cluster. Obtained with OPTICS.

Table 5.3: Table with results of the clustering efforts

descriptor	method	num clusters/eps	other	number clusters	number images clustered	mean cluster precision	mean cluster recall
pre-trained	optics	0.13	None	143	306	**1.00**	0.04
fine-tuned	dbscan	0.08	None	233	631	**1.00**	0.07
fine-tuned	dbscan	0.12	None	337	942	0.84	0.13
fine-tuned	kmeans	1700	Outlier	-	5167	0.79	0.26
fine-tuned	kmeans	2000	Outlier	-	5170	0.81	**0.27**
fine-tuned	optics	0.13	None	**791**	2015	0.89	0.15
retrain 1	dbscan	0.12	None	326	883	0.88	0.12
retrain 1	dbscan	0.1	None	490	1697	0.95	0.16
retrain 1	kmeans	2000	Outlier	-	5354	0.75	0.25
retrain 1	optics	0.13	None	806	2048	0.92	0.17
retrain 2	dbscan	0.0085	None	536	2086	0.89	0.18
retrain 2	dbscan	0.008	None	453	1531	0.95	0.16
retrain 2	dbscan	0.01	None	213	551	0.84	0.09
retrain 2	kmeans	1000	PCA	-	-	0.08	0.26
retrain 2	kmeans	2000	Outlier, PCA	-	**5866**	0.68	0.27
retrain 2	kmeans	500	Outlier, PCA	-	1531	0.93	0.16
retrain 2	optics	0.13	None	741	1924	0.93	0.16
retrain 2	optics	0.13	t-SNE	760	2017	0.90	0.14

problem is solved in OPTICS that, on its part, tends to detect outliers more generously. For this reason, we removed cluster 0 from the metrics whenever its size exceeded 100 images (considered the maximum reasonable cluster size). We observe that some efforts with DBSCAN with very low distance obtain a rather good performance (see retrain 1, dbscan, 0.1), however, we believe that the best compromise is obtained by OPTICS on retrain 2 with a distance of $\epsilon = 0.13$.

Determining which method performs best between k-means (with 2000 clusters and outlier removal) and OPTICS (with $\epsilon = 0.13$) is complex. On one side, k-means undoubtedly achieves the highest recall, while OPTICS the highest precision. We also weight the fact that k-means requires to navigate 2000 clusters, while OPTICS only slightly over 700.

If we experiment with searching for cluster groups that are also present in the morphograph, we see that k-means performs slightly better OPTICS, which has a tendency to split up the

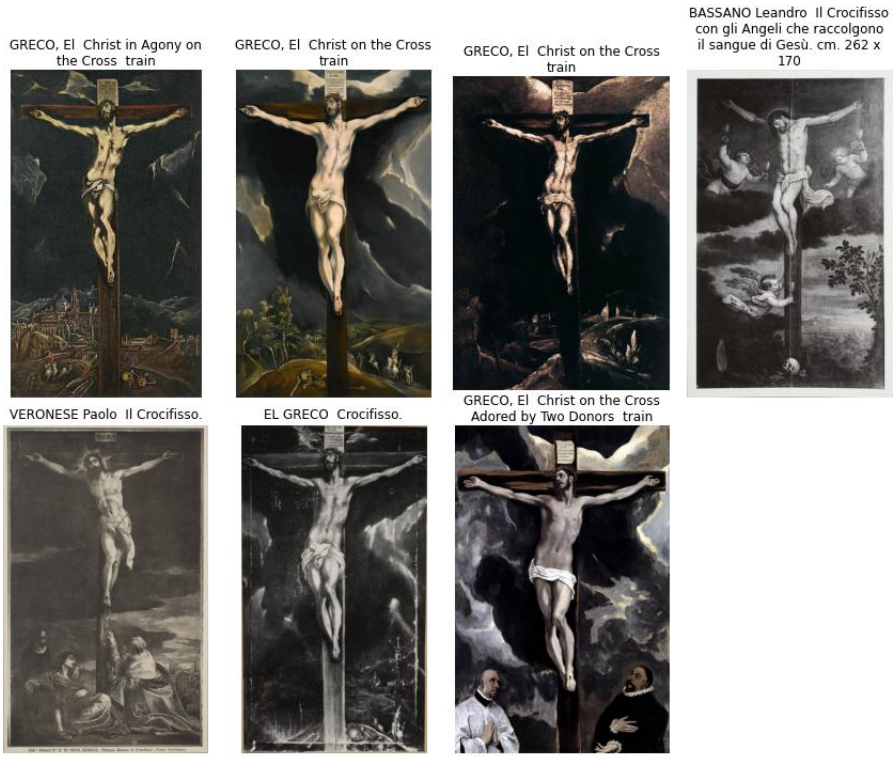


Figure 5.18: El Greco, Crucifixion cluster. Obtained with k-means.

clusters and retrieve fewer instances. In the examples reported in this thesis, Figure 5.18, 5.19 and Figure 5.16, 5.17, respectively the cluster with El Greco's Crucifixion with k-means (top) and with OPTICS (bottom) and Tiziano's Danae also with k-means and OPTICS, we see that OPTICS catches only almost exact copies with variations in the background, with an extremely high precision. K-means, on the other side, includes in the clusters also wider variations from the prototype that are, most frequently, still coherent. At times the retrieved image is too far from the original to be considered part of the cluster.

Furthermore, if we consider the nature of the clusters that contained useful additions to the morphograph, we see that k-means offers some rather interwoven and imprecise clusters, which were complex to annotate. Take in example the clusters in Figure 5.21: we see that k-means tends to agglomerate rather impure clusters. In the example on the top, the agglomeration is rather coherent in the poses (all the elements in the cluster show hands raised to the sky) but the thematic variance between the elements is too large to form a pattern; on the other side, the cluster on the bottom contains a coherent subset (the Leda and the swan pose) whilst including many impure elements.

Finally, if we select clusters at random, OPTICS is much more likely to provide a sensible cluster. The examples in Figure 5.22 highlight how the performance of OPTICS is largely superior to that of k-means on the great mole of sketches and drawings.

Figure 5.19: El Greco, Crucifixion cluster. Obtained with OPTICS.

We conclude that, for the purpose of annotating new connections, k-means with outlier removal performs best overall, with the greatest potential to find numerous and significant additions. For the final clustering, which is expected to be cleaner and essential, the potential of OPTICS is greater, as the clusters are much more coherent, thus more interesting when scrolling at random.

5.2.4 Error analysis and discussion

Taking a closer look at the final clusters produced by OPTICS will shed light on the potential and limitations of the algorithm in its current state.

After annotating the first 100 clusters, we obtain a cluster accuracy of 0.85. This is even more remarkable if we consider that the nature of the clusters annotated as wrong is predominantly drawings and architectural elements. On the other side, this puts light on some areas of better or worse performance. Alongside the already highlighted low recall in areas of high density, we observed how paintings (especially colour images) are often precisely clustered, as in the example in Figure 5.23, which shows images from the morphograph alongside two new findings (the second and fifth image from the top). On the other side, we expect sketches, furniture, architectural elements, animals, portraits and natural landscapes obtain a generally lower precision due to their much scarcer or in-existent representation in the morphograph. Interestingly, while the precision is undoubtedly lower, the model is able to pick up similarities across domains and inside remote domains. For instance, in Figure 5.24, we observe different clusters containing architectural elements. The top cluster demonstrates the ability of the model to sensibly group sketches of similar arcs into a coherent group, while the middle one demonstrates the generalisability to different settings when grouping arcs of an altar with those of doors. Finally the bottom one confirms the remarkable ability of the model to individuate the same object from different perspectives. Furthermore, we detect various cases of multi-modal cross-domain awareness or style awareness on objects. This can be seen in the examples of Figure 5.25, where, on the top, a sketch of a glass vase is juxtaposed to the real

Figure 5.20: Generic crucifixion cluster. Obtained with k-means.

object and, on the bottom, where chairs with similar manufactures were correctly grouped together. The invariance to the perspective is interestingly visible not only for architectural elements, which were to be expected due to the numerous perspective changes in the vedute by Canaletto, but also on objects, drawings and, most commonly, statues (as in Figure 5.26).

Overall, while the clusters are most often precise and the generalisability to different objects is notable, the recall of the algorithm is far too low and requires great improvement before we can detect the full range of pattern groups in the collection. In this optic, the human-in-the-loop approach does not allow to improve the recall directly as it only works on the refinement of the existing clusters. Moreover, the retraining tends to overfit even more the already overfitting model, which might decrease the generalisability and, with that, the recall. More and new clustering methods should be experimented with and a novel retraining procedure invented.

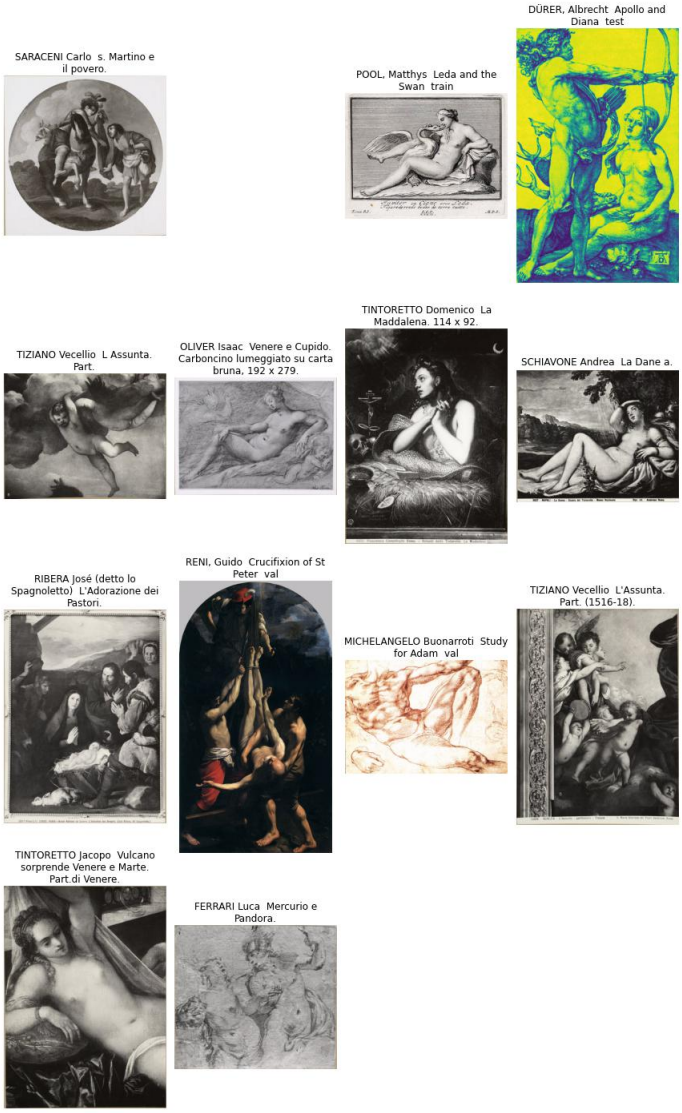


Figure 5.21: Impure Leda with Swan cluster. Obtained with k-means.

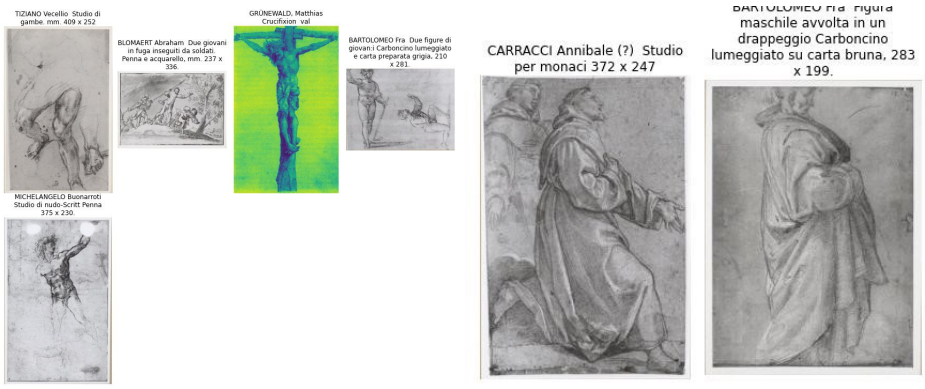


Figure 5.22: Example of clusters drawn at random. Obtained with k-means (left), and OPTICS (right).

Figure 5.23: Cluster of paintings example. Bassano, Flagellazione di Cristo. Obtained with OPTICS.

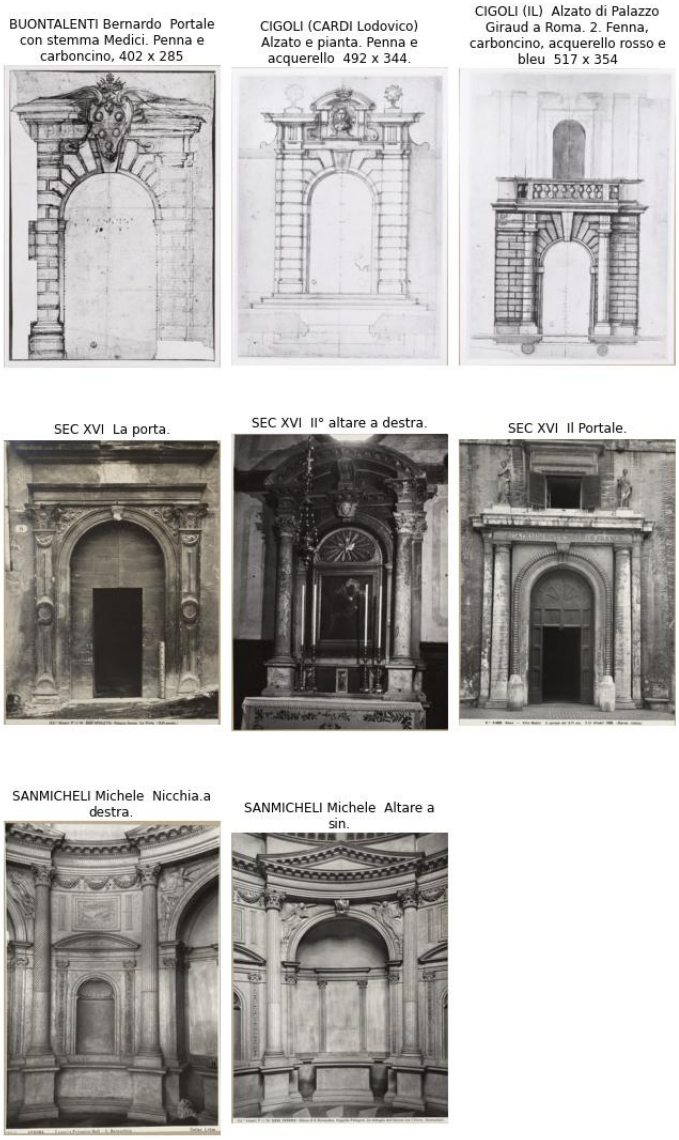


Figure 5.24: Example clusters of architectural drawings or captures. Obtained with OPTICS.

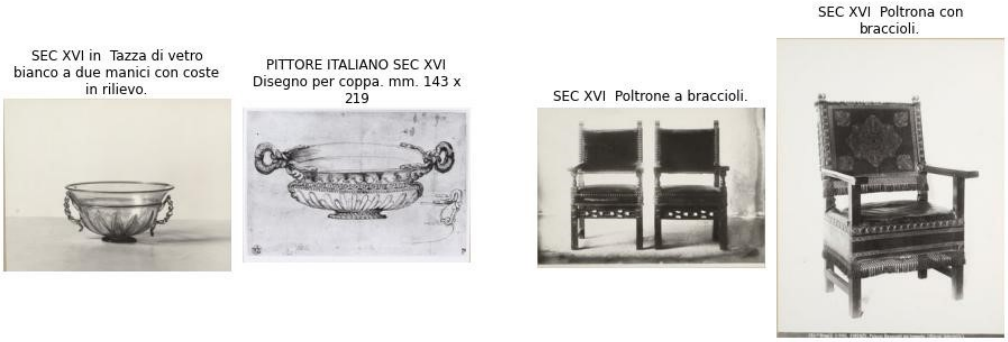


Figure 5.25: Examples of clusters of objects. A vase drawing to the object (left). Chairs (right).

Figure 5.26: Examples of clusters recognising the same object from different perspectives. Drawings of the same pose from slightly changed angles (left). Two photos of the same statue from completely different angles and scales (right)

5.3 Findings

To obtain the largest amount of possible patterns, we apply the full pipeline on the cleaned collection comprising of about 78'000 artworks and partially annotate it. We use OPTICS with $\epsilon = 0.13$ for the clustering step and obtain 2190 groups, with 4879 total clustered images, a 0.91 precision and 0.07 recall. After annotating all the aforementioned clusters, we obtain 1212 new pairs, with 749 new images, 227 of which enrich 57 existing patterns, and 522 which create 238 entirely new pattern groups. This causes an overall growth of 15.48% in the size of the morphograph.

To give an understanding of the quality of the final clusters and annotations thereof, we classify the first 200 images that were annotated as patterns into *dubious cases* and *clear patterns*. We find that about 75% of the images added exhibit a clear pattern and only 25% is dubious.

Furthermore, we divide the clusters into cases that are *similar for a machine* but share no visual influence for humans (taking a rather general perspective), *dubious cases that were not annotated* as patterns, *dubious cases that were annotated* as patterns and *clear patterns*, the last two being the aforementioned classes. We show some examples of each group to better understand the quality of the final clusters and the findings they produce.

Two examples of the clusters that were considered unrelated and were not annotated can be seen in Figure 5.27. The two clusters demonstrate how there are certain figures that look completely unrelated to any human eye, but that are tracked down by the machine for their common pose. On the other side, numerous are the cases of dubious connotation. On the one side, we show clusters that, in the end, were not annotated despite in doubt. These are visible in Figure 5.28 where we observe a considerable similarity that is not, however, indicative enough of a not coincidental connection.

On the other side, we show groups of images that were annotated as PATTERNS but are still considered dubious. This group can be seen in Figure 5.29. We added to the group on the top the third image from the left. We observe an uncanny similarity between the various images, the poses coincide almost exactly. However, there are a few relevant differences: the crucifixion is a representation with very few degrees of freedom and a change in the drape may be decisive. The two women share the same representational standard but are possibly different models. Pontormo's Nude, despire resembling very closely the bust of Michelangelo's Adamo, has a different arm position.

Finally, we show some examples of clear patterns that were found or enriched in this thesis in Figure 5.30. The similarity inside each of the groups is immediately striking. Sassoferrato's Madonna most probably the same woman and in the same pose, with only minor changes in the clothing and lighting. We immediately understand that for Sassoferrato to have painted the second Madonna, he must have been looking at the first. While the first case is an autograph pattern, the following cases show examples of complex attribution (Pittore Sec.) who were likely to have taken inspiration from the original work. This is the case, we believe, with

Figure 5.27: Examples of clusters proposed by the model that do not appear related to a human eye. On the left, we observe a male passive figure with a dangling arm being connected to Venus, holding on her arm. On the right, we see once again a male and a female figure being grouped on the base of their posture and the arm.

Tintoretto's Deposition. Additionally, Raffaello's Madonna with Child, that was already in the morphograph, has now been traced in a work attributed to Ghirlandaio, who clearly took inspiration in the pose of the Child from the central Madonna. Finally, the second cluster from the top shows three drawing of the same statue, Michelangelo's Sansone, which were probably realised in the presence of the statue but from different angles.

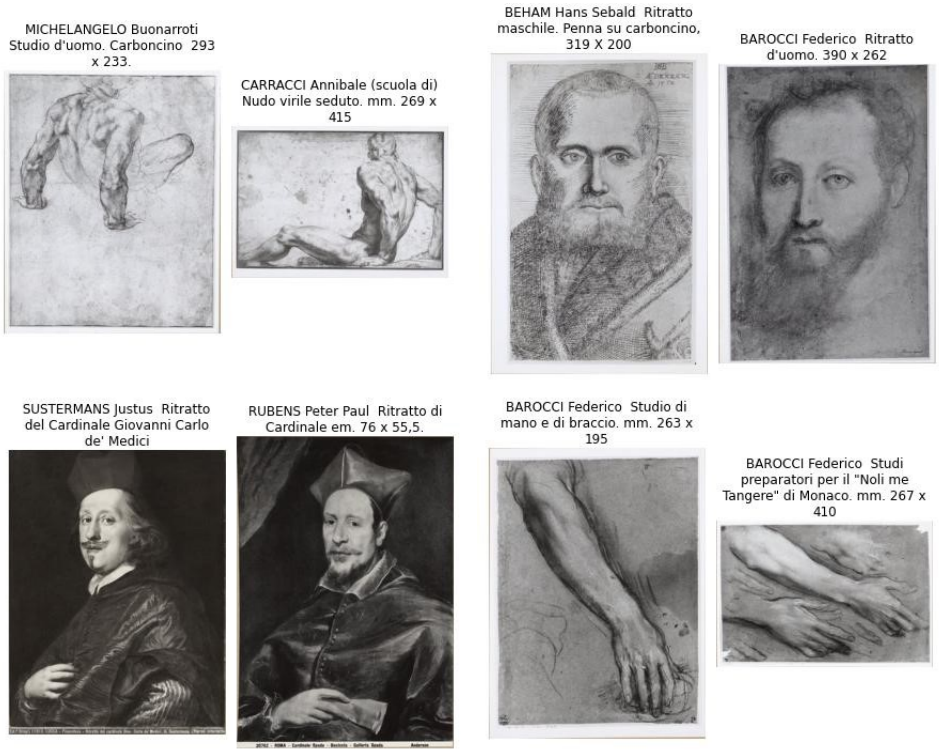


Figure 5.28: Examples of clusters proposed by the model that share some clear features but were not annotated as patterns (connections that were annotated as SIMILAR instead of POSITIVE). From top to bottom left to right, we observe two seated male figures whose similarity may be coincidental. We observe two portraits with extremely similar eyes but whose relation is almost merely stylistic. Two cardinals in the same dress in different poses and perspectives. Hand studies.

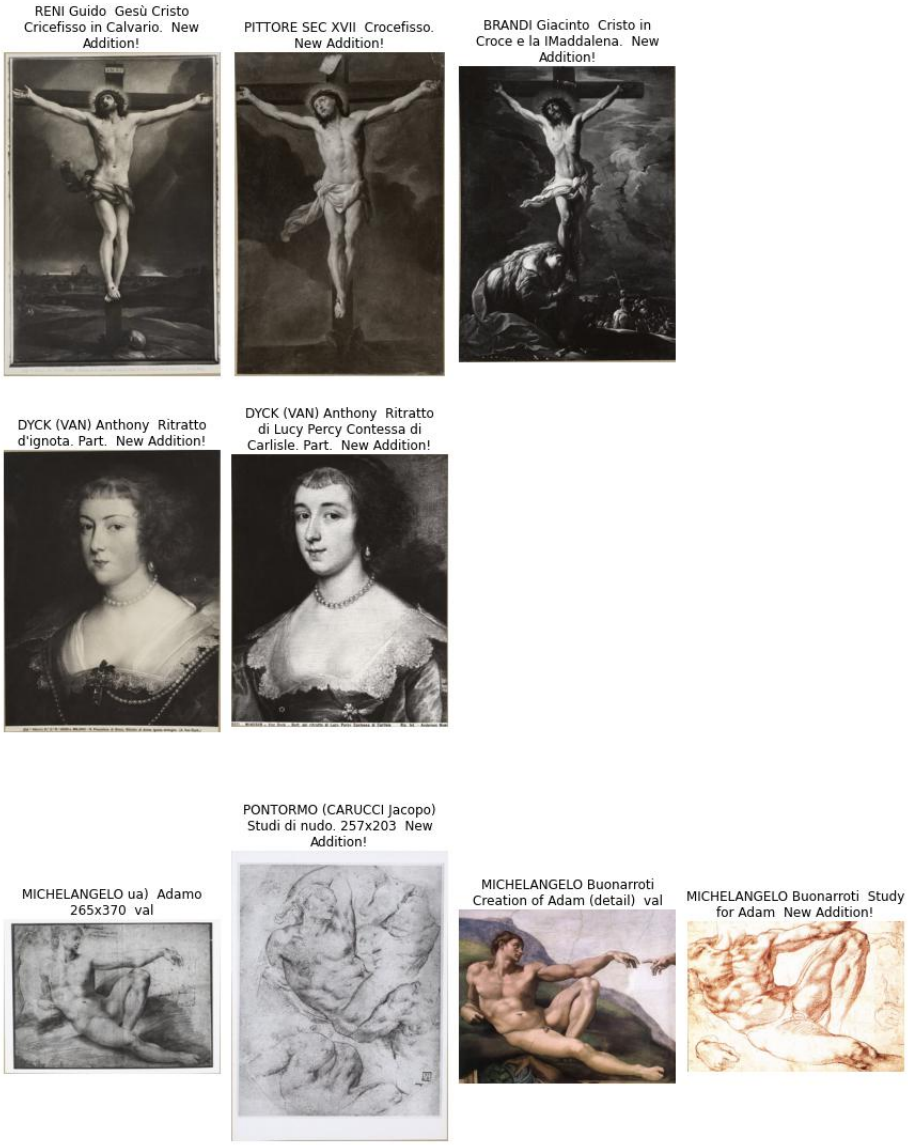


Figure 5.29: Examples of clusters proposed by the model that were annotated as patterns but are still considered dubious cases. When an image was added to the morphograph as part of this thesis it is indicated by 'New Addition!', when it was already present, we indicate the set it belonged to.

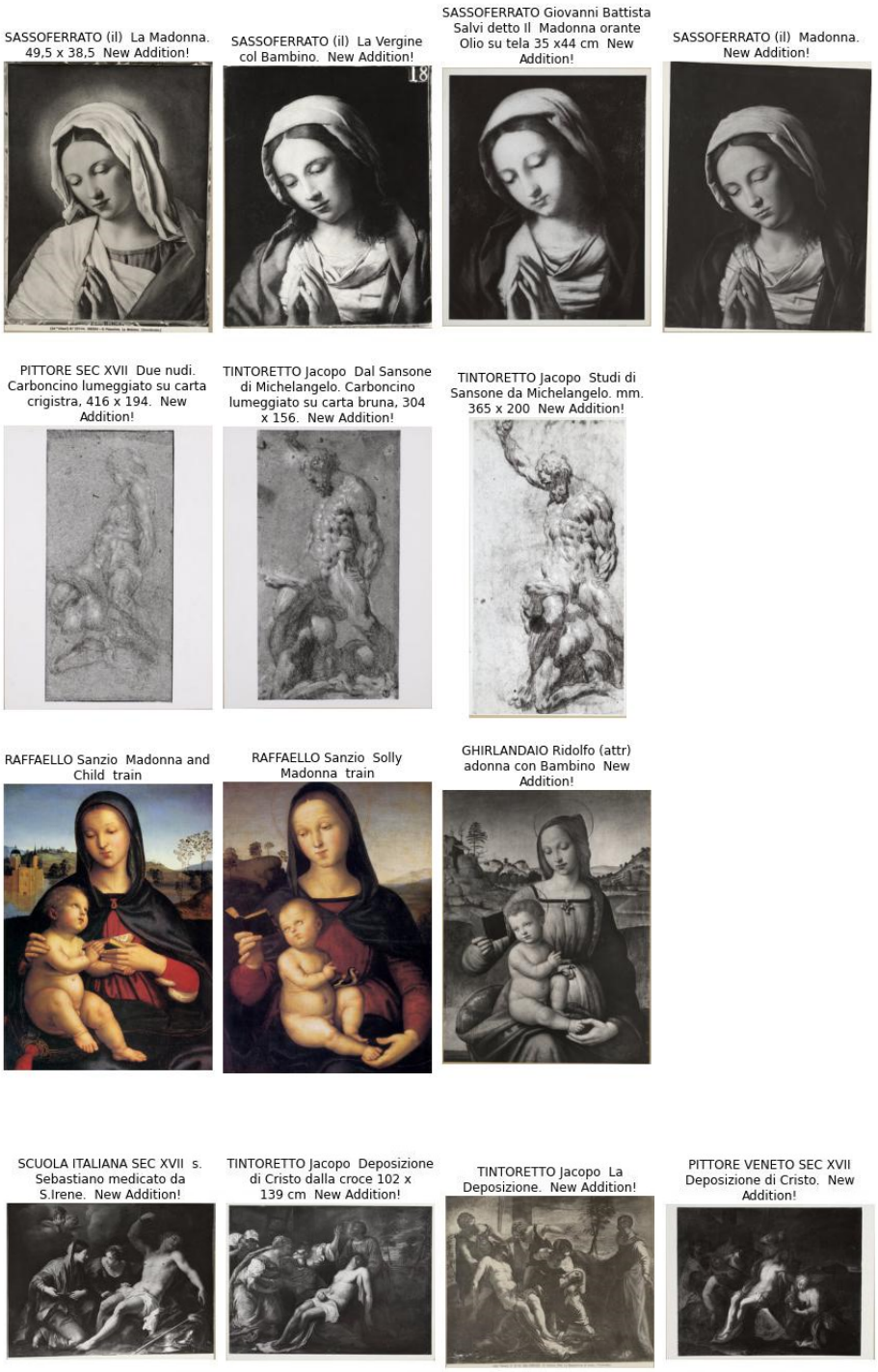


Figure 5.30: Examples of clusters proposed by the model that were considered as clear patterns.

6 Discussion

6.1 What has been done, what has not and what cannot

We believe that the contribution of this thesis lies on the elimination of the middle man of the input image, which was essential in previous attempts (B. Seguin, 2018; Shen et al., 2019). Such an elimination shifts the focus from a research based on particular intuitions of researchers, who query the retrieval model with an input image to investigate its patterns, to an unbiased suggestion of patterns. The first is based on the preemptive knowledge that the searched image can yield interesting results and, in turn, the results are based on the suggestions of the art historian.

In our effort, however, removing the need for an input image, we increase the scalability and the breadth of such findings, automatically generating spontaneous suggestions (the patterns) and annotating the correct ones. We, therefore, remove the bias of the researcher, who no longer finds only what they are looking for, but everything that has been suggested automatically.

At the same time, this approach does not completely eliminate bias. A model learns from what it sees. In our case, the model sees the morphograph and bases its representation of patterns from it. Immediately, we detect a number of biases which the morphograph has exposed the model to. We mentioned in the Results how, due to the highly imbalanced number of examples in the training set with higher attention being devoted to the Child than the Madonna, this is mirrored in the limitations of the findings. Furthermore, the morphograph contains examples of patterns that occupy either the full scale of the image or a considerable percentage of it. For this reason, the model could not learn satisfactorily to detect small pattern repetitions. While we believe that considerable improvements can be achieved towards a complete scale invariance and detail recognition using object detection methods on the ground-truth and cropping around the regions of interest; a thorough annotation of examples containing detail patterns is necessary. Even more so, the morphograph is limited to patterns in the domains of paintings, engravings and sketches. While we demonstrated a certain generalisability of the model, partly due to the pre-training, we believe that the groundtruth should be enriched

with architectural and design patterns but also statues to obtain a large scale recognition of patterns.

Overall, we observed in the Results section how the semi-supervised model could successfully detect patterns despite the existing barriers of style, medium, scale, deterioration, perspective, and mirroring. This is in line with the results that had already been obtained by B. Seguin, 2018. Along the same lines as B. Seguin, 2018, we could also observe that our methods could not obtain the desired invariance to content. We question whether this task is entirely apt for CNN networks. In fact, the pooling layer almost completely removes any spatial and superficial information, maintaining only the compressed, learned, representation. Perhaps hybrid methods, such as the spatial re-ranking[I] explained in B. Seguin, 2018, might prove to be most suited. Nonetheless, these methods currently suffer from two issues: terrible scalability and absence of a compact descriptor to be used efficiently for clustering, bypassing the need for a similarity matrix, and should therefore be further investigated.

The performance of the model was nonetheless excellent and we achieved a clustering quality and 2D spatial representation of the images of unprecedented precision. If we compare our results with those achieved by Castellano et al., 2021; Diagne and Barradeau, n.d.; Duhaime, 2019; Saleh et al., 2016b, we observe a much finer grained representation of the clusters and the space. The step is promising and a considerable amount of new annotations are already available for art historical studies to validate their connection and broaden the findings on pattern propagation.

6.2 Personal reflections

Writing a thesis in Digital Humanities may sometimes feel like walking on a rope. It comes with the awareness that developing methods for art historians should not be blind to art history and should stem from a comprehension of the art historical task, of its motivation and of its application, and an intuition of the quality of the results. At the same time, it should also be technically advanced, because a technically trained person should bring this value to the table. I admit, however, that I have fallen from the rope a few times.

I started this thesis from B. Seguin, 2018's work: I had the data, the morphograph and an old uncommented code that needed to be re-written completely to a new library and updated. So I dove head first into the code and coded for three months, trying uncountable different configurations, parameters, adding new features, new methods, clustering, and making the platform. It was only after I made the platform, roughly two months and a half into the thesis, that I started really looking at the morphograph. I had seen some examples of groups in the morphograph before, but somehow I always ended up on the same ones. It was only then that I understood the real difficulties of the task I had in hand. I gained an understanding of the amount of clusters we could aim for, of their sizes, their purity, the varied nature of

[I] After some experimentation, we decided not use the spatial re-ranking in this thesis, as it required too much memory and compute power and it does not allow to create a compressed representation of the image.

the morphograph, but also of the surprisingly large number of photos of chairs in the data. Even worse, I started asking myself what was meant with patterns about three weeks before the submission of this thesis. I googled it, as any respectable computer scientist, and found nothing. I felt like a machine, I had also constructed my understanding of patterns from the examples I saw.

This made me wonder what someone with a technical background can give to Digital Humanities. I came to reflecting about my own experience and the following is what I could notice:

- The field of deep learning is rapidly changing and a deep learning researcher has be dedicated to being on top of this wave, so as not to drown under it.
- Deep learning, for as much as we would like to believe oppositely, is very much not a science in the practice, it is trial and error field where intuition is the main motor. Great familiarity with the methods is the only way to navigate the field.
- Due to the black box nature of neural networks, we are forced to treat the models as a behavioural psychologist would treat their patient: we need to gain an understanding of how its mind works but can only interpret its actions. A profound understanding of the mathematical and procedural steps, as well as practice, are essential for assessing the model decisions.
- As digital humanists, we need to be able to flexibly switch from one field to another, acknowledging their differences in research style, writing, thinking, and nature of findings.
- We should not isolate from the people who would benefit from the findings. We do not know what they really want unless we ask.

These points are by no means exhaustive nor written in stone, they are but personal reflections.

7 Conclusion

To the best of our knowledge, we carried out the first attempt at semi-supervised clustering of patterns in art history. We shifted the attention from previous search-engine-based results to a global view organised in clusters (in a 2D continuous space). While previous efforts focused on image retrieval for artistic patterns (B. Seguin, 2018; Shen et al., 2019), which is an input based discovery system, we eliminated the intermediary of the input image, obtaining an organic view of the patterns. We believe that this attempt took a, minor, step towards a large-scale comprehensive discovery of visual patterns in the history of art.

In this endeavour, we set up an integrated end-to-end pipeline to learn task specific feature descriptors of artworks. We cluster and evaluate the descriptors, annotate the produced clusters expanding the original morphograph, translate the annotations into a training set for iterative learning, and retrain. We completed the loop three times (including the first training) and presented the final results in a threefold manner: we present the enlarged morphograph, the final clusters of almost 80'000 artworks and integrate the clusters in a 2D cartography of propagation of images, created from a projection of the learned descriptor space (using `pixplot`).

For the feature learning step, we proposed a number of variations over the original triplet learning approach of B. Seguin, 2018 and demonstrated their effectiveness on our task. In particular, we adopted the recent model of ResNeXt, which introduces the cardinality dimension to the neural architecture with remarkable improvements over the base architecture of ResNet (Xie et al., 2017. We forged a new, compound, loss that enriches the Hinge margin loss with an anchor swap and positive distance upper-bound (Balntas et al., 2016; Ho et al., 2021. We showed a significant improvement from the pre-trained network, increasing the recall at 20 from 0.19 to 0.5 and the mean average precision from 0.13 to 0.41.

In the feature clustering, we evaluated the performance of the different methods and pipelines for the different tasks in this thesis. On one side, we demonstrated the suitability of our novel compound method of k-means with DBSCAN outlier removal for the annotation task, that posits greater importance on the recall in the precision-recall trade-off. We further observed

the appropriateness of OPTICS with our positive distance upper-bound as max_eps for the final clusters visualisation, where precision is key.

Moreover, given the semi-supervised approach to the clustering, we coined some ad-hoc morphograph-based precision and recall metrics, which capture, respectively, the purity of the clusters containing morphograph images and the proportion of morphograph groups that are captured to some extent in the clusters. We also add a supervised metric of cluster accuracy based on a human annotation of the clusters. With the designed metrics, we showed the remarkable final precision of 0.93 and the rather low recall of 0.16, which, however, was a considerable improvement over the 0.04 recall on the pre-trained descriptors.

To take full advantage of the triplet learning, we devised a tailored annotation interface which allows to annotate both correct patterns inside clusters and wrong images that need to be removed from the clusters. We use the positive connections to enrich the morphograph and the negative connections to be added to the hard negative sampling. This focuses the learning on the most problematic clusters and enriches the learning with the new findings. Furthermore, in order to avoid excessive overfitting and, at the same time, preventing the model from diverging, we retrained for a 'half' set, including half the training triplets and the added negative and positive connections. After every step, we include an average of 100 new images to the morphograph, showing the usefulness of the iterative approach.

Finally, we executed the full pipeline on a larger set of about ten times the original size (from 8'900 to almost 80'000) and annotated a part of the clusters. We found over 700 new images that enriched the morphograph from the original size of 1800 images.

We served all the results on an online interface accessible here. In the page morphograph, we present the ground-truth set with its new additions in red. Given the growing size of the set (now containing over 400 clusters), we introduced a number of sorting options that are intended to address the possible avenues of research for art historians. This step is only a first, exploratory, avenue of research, which we believe will become crucial for future developments and additions to the morphograph. Additionally, we included a clusters viewer page which allows to navigate the final clusters and query using metadata textual search. Finally, we re-implemented the `pixplot` 2D scalar spatial image visualisation using the t-SNE projection of our feature descriptors. Although not explored in the results, we believe that the spatial visualisation with the cluster search option can yield interesting outlets for research.

Overall, we believe this research paved the way for numerous interesting avenues of. Particularly, we believe that work should be dedicated to the scalability of the pipeline to more and larger datasets in order to be able to map the full set of digitised artworks available online. Methodologically speaking, more attention should be devoted to avoiding overfitting and increasing cluster recall. Finally, given the increasing size of the morphograph, the raised issues of navigability should be addressed, perhaps investigating novel visualisation path, such as a CAVE or Virtual Reality experience.

On a more general note, we wish to obtain more trustworthy metadata, to explore hybrid text-image approaches and improve the original recognition.

Bibliography

Alani, H., Dijkshoorn, C., Jongma, L., Aroyo, L., van Ossenbruggen, J., Schreiber, G., ter Weele, W., & Wielemaker, J. (2018). The rijksmuseum collection as linked data. *Semant. Web*, *9*(2), 221–230. https://doi.org/10.3233/SW-170257

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). Optics: ordering points to identify the clustering structure. *ACM Sigmod record*, *28*(2), 49–60.

Babenko, A., Slesarev, A., Chigorin, A., & Lempitsky, V. (2014). Neural codes for image retrieval. *European conference on computer vision*, 584–599.

Balntas, V., Riba, E., Ponsa, D., & Mikolajczyk, K. (2016). Learning local feature descriptors with triplets and shallow convolutional neural networks. *Procedings of the British Machine Vision Conference 2016*, 119.1–119.11. https://doi.org/10.5244/C.30.119

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Capelli, S. (n.d.). Le copie pittoriche della pietà di michelangelo per vittoria colonna: marcello venusti e copisti anonimi. attribuzioni e precisazioni. Retrieved June 19, 2022, from https://www.academia.edu/12724495/Le_copie_pittoriche_della_Piet%C3%A0_di_Michelangelo_per_Vittoria_Colonna_Marcello_Venusti_e_copisti_anonimi_Attribuzioni_e_precisazioni

Castellano, G., Lella, E., & Vessio, G. (2021). Visual link retrieval and knowledge discovery in painting datasets. *Multimedia Tools and Applications*, *80*(5), 6599–6616. https://doi.org/10.1007/s11042-020-09995-z

Castellano, G., & Vessio, G. (2021). Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview. *Neural Computing and Applications*, *33*(19), 12263–12282. https://doi.org/10.1007/s00521-021-05893-z

Castelnuovo, E., Ginzburg, C., & Curie, M. (2009). Symbolic domination and artistic geography in italian art history. *Art in Translation*, *1*(1), 5–48. https://doi.org/10.2752/175613109787307672

Cetinic, E., Lipic, T., & Grgic, S. (2018). Fine-tuning convolutional neural networks for fine art classification. *Expert Systems with Applications*, *114*, 107–118. https://doi.org/10.1016/j.eswa.2018.07.026

Dal Pozzolo, E. M. (2006). 'la 'bottega'di tiziano: sistema solare e buco nero." *Studi tizianeschi*, *4*, 53–98.

Das, D., Ghosh, R., & Bhowmick, B. (2019). Deep representation learning characterized by inter-class separation for image clustering. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 628–637.

Dempsey, C. (1980). Some observations on the education of artists in florence and bologna during the later sixteenth century. *The Art Bulletin*, *62*(4), 552–569. https://doi.org/https://doi.org/10.2307/3050053

Deng, Y., Tang, F., Dong, W., Ma, C., Huang, F., Deussen, O., & Xu, C. (2021). Exploring the representativity of art paintings [Conference Name: IEEE Transactions on Multimedia]. *IEEE Transactions on Multimedia*, *23*, 2794–2805. https://doi.org/10.1109/TMM.2020.3016887

Diagne, C., & Barradeau, N. (n.d.). *Google art t-sne*. https://artsexperiments.withgoogle.com/tsnemap/

Didi-Huberman, G. (1996). Pour une anthropologie des singularités formelles. remarque sur l'invention warburgienne [Publisher: Editions Belin]. *Genèses*, (24), 145–163. Retrieved June 15, 2022, from https://www.jstor.org/stable/26201510

Diers, M., Girst, T., & von Moltke, D. (1995). Warburg and the warburgian tradition of cultural history [Publisher: [New German Critique, Duke University Press]]. *New German Critique*, (65), 59–73. https://doi.org/10.2307/488533

di Lenardo, I., Seguin, B. L. A., & Kaplan, F. (Eds.). (2016). *Visual patterns discovery in large databases of paintings* [Meeting Name: Digital Humanities 2016].

Duhaime, D. (2019). *Pixplot*. https://github.com/YaleDHLab/pix-plot

Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., & Mazzone, M. (2018). The shape of art history in the eyes of the machine. *9*(1).

Ericani, G., Caramanna, C., & Millozzi, F. (Eds.). (2010). *Jacopo Bassano, i figli, la scuola, l'eredità. 1*. Museo civico.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, *96*(34), 226–231.

Etro, F., & Pagani, L. (2013). The market for paintings in the venetian republic from renaissance to rococò [Publisher: Springer]. *Journal of Cultural Economics*, *37*(4), 391–415. Retrieved June 26, 2022, from https://www.jstor.org/stable/43549853

Gombrich, E. H. (1961). *Art and illusion*. Pantheon Books New York.

Gombrich, E. H. (1965). The style "all 'antica": imitation and assimilation [Reprinted in Norm and Form, 1966]. *Studies in Western Art*, 31–41.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. *MIT press*.

Gordo, A., Almazán, J., Revaud, J., & Larlus, D. (2017). End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, *124*(2), 237–254. https://doi.org/10.1007/s11263-017-1016-8

Guichard, C. (2010). Du «nouveau connoisseurship» à l'histoire de l'art original et autographie en peinture. *Annales. Histoire, Sciences Sociales*, *65*(6), 1387–1401.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (1), 770–778. https://doi.org/10.1109/CVPR.2016.90

He, T., Wei, Y., Liu, Z., Qing, G., & Zhang, D. (2018). Content based image retrieval method based on SIFT feature. *2018 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*, 649–652. https://doi.org/10.1109/ICITBS.2018.00169

Ho, K., Keuper, J., Pfreundt, F.-J., & Keuper, M. (2021). Learning embeddings for image clustering: an empirical study of triplet loss approaches [ISSN: 1051-4651]. *2020 25th International Conference on Pattern Recognition (ICPR)*, 87–94. https://doi.org/10.1109/ICPR48806.2021.9412602

Jahrer, M., Grabner, M., & Bischof, H. (2008). Learned local descriptors for recognition and matching. *Computer Vision Winter Workshop*, *2*(3), 103–118.

Jost, F. (1966). Imitation. *Proceedings ICLA*, 695–936.

Kanagala, H. K., & Jaya Rama Krishnaiah, V. (2016). A comparative study of k-means, DBSCAN and OPTICS. *2016 International Conference on Computer Communication and Informatics (ICCCI)*, 1–6. https://doi.org/10.1109/ICCCI.2016.7479923

Kaufmann, T. D. (2004). *Toward a geography of art*. University of Chicago Press.

Kaufmann, T. D., Dossin, C., Joyeux-Prunel, B., & Kaufmann, T. D. (Eds.). (2015). *Circulations in the global history of art*. Ashgate.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Kultermann, U. (1990). Woman asleep and the artist [Publisher: IRSA s.c.]. *Artibus et Historiae*, *11*(22), 129–161. https://doi.org/10.2307/1483403

La Malfa, C. (2000). The chapel of san girolamo in santa maria del popolo in rome. new evidence for the discovery of the domus aurea. *Journal of the Warburg and Courtauld Institutes*, *63*(1), 259–270.

Lecoutre, A., Negrevergne, B., & Yger, F. (2017). Recognizing art style automatically in painting with deep learning. *16*(1).

Loh, M. H. (2007). *Titian remade. repetition and the transformation of early modern italian art*. Getty Publications.

MacQueen, J. (1967). Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*, 281–297.

Martin, J. R. (1968). Two terra-cotta replicas of the laocoon group. *Record of the Art Museum, Princeton University*, *27*(2), 68–71. Retrieved June 28, 2022, from http://www.jstor.org/stable/3774488

Mukherjee, S., Asnani, H., Lin, E., & Kannan, S. (2019). Clustergan: latent space clustering in generative adversarial networks. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 4610–4617.

Muller, J. M. (1982). Rubens's theory and practice of the imitation of art. *The Art Bulletin*, *64*(2), 229–247.

Nina, O., Moody, J., & Milligan, C. (2019). A decoder-free approach for unsupervised clustering and manifold learning with random triplet mining. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Panofsky, E. (2018). *Studies in iconology: humanistic themes in the art of the renaissance*. Routledge.

Patrick, M. ((2014). *Connoisseurship, l'œil, la raison et l'instrument*. actes de colloque à l'École du Louvre, 20, 21 et 22 octobre 2011, Paris, Éd. du Louvre.

Pozzolo, E. M. D. (n.d.). Per la storia delle falsificazioni di el greco: una prima categorizzazione con qualche esempio, in "artibus et historiae", XXXVII, 2016, 72, pp. 153-173. Retrieved June 19, 2022, from https://www.academia.edu/29784465/Per_la_storia_delle_falsificazioni_di_El_Greco_una_prima_categorizzazione_con_qualche_esempio_in_Artibus_et_historiae_XXXVII_2016_72_pp_153_173

Prasad, V., Das, D., & Bhowmick, B. (2020). Variational clustering: leveraging variational autoencoders for image clustering [ISSN: 2161-4407]. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–10. https://doi.org/10.1109/IJCNN48605.2020.9207523

Racco, T. (2016). Darkness in a positive light: negative theology in caravaggio's conversion of saint paul. *Artibus et Historiae*, (73), 285.

Radenović, F., Tolias, G., & Chum, O. (2018). Fine-tuning CNN image retrieval with no human annotation [version: 2]. *arXiv:1711.02512 [cs]*. Retrieved March 15, 2022, from http://arxiv.org/abs/1711.02512

Sabatelli, M., Kestemont, M., Daelemans, W., & Geurts, P. (2019). Deep transfer learning for art classification problems. *L. Leal-Taixé & S. Roth (Eds.), Computer Vision – ECCV 2018 Workshops*, *11130*(1), 631–646. https://doi.org/10.1007/978-3-030-11012-3_48

Saleh, B., Abe, K., Arora, R. S., & Elgammal, A. (2016a). Toward automated discovery of artistic influence. *Multimedia Tools and Applications*, *75(7)*(1), 3565–3591. https://doi.org/10.1007/s11042-014-2193-x

Saleh, B., Abe, K., Arora, R. S., & Elgammal, A. (2016b). Toward automated discovery of artistic influence. *Multimedia Tools and Applications*, *75*(7), 3565–3591. https://doi.org/10.1007/s11042-014-2193-x

Seguin, B. (2018). The replica project: building a visual search engine for art historians. *XRDS: Crossroads, The ACM Magazine for Students*, *24*(3), 24–29. https://doi.org/10.1145/3186653

Seguin, B. (2019, February 28). *Replica core* [original-date: 2017-07-24T16:07:31Z]. Retrieved February 23, 2022, from https://github.com/SeguinBe/Replica-Core

Seguin, B., Striolo, C., diLenardo, I., & Kaplan, F. (2016). Visual link retrieval in a database of paintings. In G. Hua & H. Jégou (Eds.), *Computer vision – ECCV 2016 workshops* (pp. 753–767). Springer International Publishing. https://doi.org/10.1007/978-3-319-46604-0_52

Seguin, B. L. A. (2018). *Making large art historical photo archives searchable* (Doctoral dissertation). EPFL. Lausanne. https://doi.org/10.5075/epfl-thesis-8857

Shen, X., Efros, A. A., & Aubry, M. (2019). Discovering visual patterns in art collections with spatially-consistent feature learning, 9278–9287. Retrieved April 21, 2022, from https://openaccess.thecvf.com/content_CVPR_2019/html/Shen_Discovering_Visual_Patterns_in_Art_Collections_With_Spatially-Consistent_Feature_Learning_CVPR_2019_paper.html

Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., & Moreno-Noguer, F. (2015). Discriminative learning of deep convolutional feature point descriptors. *Proceedings of the IEEE international conference on computer vision*, 118–126.

Song, C., Liu, F., Huang, Y., Wang, L., & Tan, T. (2013). Auto-encoder based data clustering. *Iberoamerican congress on pattern recognition*, 117–124.

Spear, R. E. (2002). Di sua mano. *Memoirs of the American Academy in Rome. Supplementary Volumes, 1*, 79–98. Retrieved June 30, 2022, from http://www.jstor.org/stable/4238447

Tagliaferro, G., Aikema, B., Mancini, M., Martin, A. J., & Vecchi, T. (2009). *Le botteghe di tiziano*. Alinari 24 ore.

Tan, M., & Le, Q. V. (2020, September 11). EfficientNet: rethinking model scaling for convolutional neural networks [Number: arXiv:1905.11946]. Retrieved June 21, 2022, from http://arxiv.org/abs/1905.11946

Tolias, G., Sicre, R., & Jégou, H. (2016). Particular object retrieval with integral max-pooling of CNN activations. *arXiv:1511.05879 [cs]*. Retrieved April 26, 2022, from http://arxiv.org/abs/1511.05879

Van Noord, N., Hendriks, E., & Postma, E. (2015). Toward discovery of the artist's style: learning to recognize artists by their artworks. *IEEE Signal Processing Magazine, 32*(4), 46–54.

Vander Auwera, J. (2007). Rubens and his visual sources. In J. vander Auwera et al. (Ed.), *Rubens: a genius at work: the works of peter paul rubens in the royal museums of fine arts of belgium reconsidered* (p. 66). Lannoo Uitgever.

Vasari, G. ([1550] 1876II). *Le vite dei più eccellenti pittori, scultori e architetti* (Vol. 1). Guttemberg.

Warners, J. (1956). Translatio-imitatio-aemulatio. *De nieuwe taalgids 49* (pp. 289–295).

Wittkower, R. (1965). *Imitation, eclecticism, and genius*. Johns Hopkins Press.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017, April 10). *Aggregated residual transformations for deep neural networks* (arXiv:1611.05431) [version: 2 type: article]. arXiv. Retrieved June 8, 2022, from http://arxiv.org/abs/1611.05431